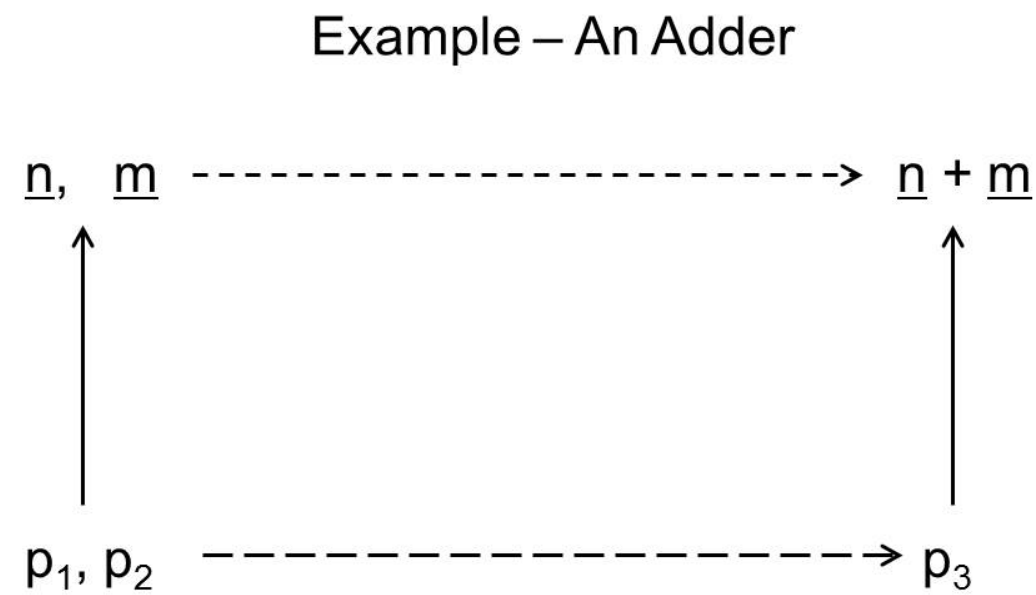
How to Think about Mental Content

Frances Egan
Rutgers University

1. Introduction: Representationalism

Most theorists of cognition endorse some version of *representationalism*, which I will understand as the view that the human mind is an *information-using* system, and that human cognitive capacities are representational capacities. Of course, notions such as 'representation' and 'information-using' are terms of art that require explication. As a first pass, representations are "mediating states of an intelligent system that carry information" (Markman & Deitrich, 2000, 471). They have two important features: (1) they are *physically realized*, and so have causal powers; (2) they are *intentional*, in other words, they have meaning or representational content. This presumes a distinction between a representational *vehicle* – a physical state or structure that has causal powers and is responsible for producing behavior – and its *content*. Consider the following characterization of a device that computes the addition function[1]:

[1] Readers will recognize the similarity to Cummins (1989) 'tower-bridge' idea.



A physical system computes the addition function just in case there exists a mapping from physical state types to numbers, such that physical state types related by a causal state-transition relation (p_1, p_2, p_3) are mapped to numbers n, m, and n+m related as addends and sums. Whenever the system goes into the physical state specified under the mapping as n, and then goes into the physical state specified under the mapping as m, it is caused to go into the physical state specified under the mapping as n+m. The bottom span depicts the representational *vehicles* and the top span their *contents*.[2]

[2] Actually, this picture is an oversimplification: the bottom level is a compression of several "implementation" levels, because representational vehicles are not physical state types. Characterizing them – say, as symbols, or nodes in a network – involves significant abstraction and idealization.

There is significant controversy about what can legitimately count as a representation.[3] The issue concerns both vehicle and content. In this paper I will focus on content. Recall Markman & Dietrich's definition – representations are "mediating states of an intelligent system that carry information." Our question, then, is 'how is the notion of *carrying information* to be understood?' The depiction of the adder, and the account of computation that it presumes, provide a straightforward answer: a state carries information just in case it is assigned a content by a mapping or interpretation function. But we want something more than this. We want to understand how the assignment of content plays a role in the *explanation* of mentality, if indeed it does. Accordingly, my concern in this paper is with the cognitive sciences that attempt to explain our representational capacities -- in particular, with computational cognitive psychology and computational neuroscience – and not with fields that presume our representational capacities, and so may be seen as continuous with commonsense. First, I consider two proposals for how the central notion of 'representation' is understood in computational cognitive science. Both accounts, I argue, misconstrue the role of representation in computational models. In developing an alternative proposal I identify and characterize *two* kinds of content – *mathematical content* and *cognitive content* – that play distinct roles in computational cognitive theorizing. I conclude by considering the prospects of computational cognitive science for explaining *intrinsic intentionality*.

2. <u>The Received View: *Hyper Representationalism*</u>

Jerry Fodor has recently made the following claim:

> Cognitive science consists mostly of working out the distinction between *represent-*

---

[3] See, for example, Ramsey 2007.

> *ing as* and merely *representing*, which is to say that cognitive science is mostly about intentional states *as such*. (2008, 13)

The claim that 'cognitive science is… about intentional states *as such*' can be understood as the thesis that cognitive theories make explanatory appeal to states that are *essentially* intentional. A state is essentially intentional if and only if it not only *has* content, but whatever content it has it has *essentially* – necessarily, if it had a different content, or no content at all, it would be a different *type* of state.[4] This is a crucial component of the received view of mental representation.

The robust notion that Fodor has in mind, the notion that contrasts with "merely representing", allows for the possibility of *mis*representation. Intentional states *as such* differ from things that are said to have (Gricean) 'natural meaning' – such things as the rings in a tree's trunk and smoke, which represent the tree's age and the presence of fire respectively – in that they can, and occasionally do, misrepresent.[5]

The received view also requires that for a mental state or structure to genuinely represent an object or property some *naturalistic* relation must hold between the two. The "tower-bridge" notion requires only a mapping between the representing vehicles and the represented objects or properties, but, of course, mappings are cheap. It is thought that some more

---

[4] A property (or set of properties) is *essential* only relative to a particular taxonomy or way of type-individuating. So the relevant claim is that cognitive science type-individuates mental states in such a way as their contents are essential.

[5] Fred Dretske defends the significance of misrepresentation as follows:

> It is the power to misrepresent, the capacity to get things wrong, to say things that are not true, that helps *define* the notion of interest. That is why it is important to stress a system's capacity for misrepresentation. For it is only if the system has this capacity does it have, in its power to get things right, something approximating *meaning*. (1988, 65)

robust relation is required to justify the claim that items in the lower span actually represent items in the upper span. The relevant relation might be "information-theoretic" (based on causal relations) or *teleological* (e.g. based on evolutionary function), or some other relation that is specifiable in non-intentional and non-semantic terms.[6] So-called 'naturalistic psychosemantics' is the attempt to specify the robust content-determining relation that holds between a mental state and the object, property, or state of affairs it is about.[7]

Why must the content-determining relation be *naturalistic*? Proponents of the received view have something like the following in mind: only if the representation relation is, at least in principle, specifiable in non-semantic and non-intentional terms will computational cognitive science deliver on its promise to provide a fully mechanical account of the mind, and provide the basis for a naturalistic account not only of cognitive capacities, but also of intentionality. The idea is that intentional mental states are essentially so, but intentionality is not a primitive property of the natural world. It is this promise that accounts for much of the interest among philosophers of mind in computational cognitive science, since it seems to promise a naturalistic reduction of intentionality.

So summing up the 'received view': the relevant notion of representation in cognitive science requires (1) that mental representations have their contents essentially, (2) that misrepresentation is possible, and (3) that such content is determined by a privileged naturalistic

---

[6] Of course, the content-determining relation must allow for *mis*representation, so there must, in principle, be some circumstances where the specified relation holds between the internal state or structure and some other object or property. Naturalistic theories often founder trying to satisfy this requirement.

[7] Proponents of naturalistic psychosemantics include Dretske (1981, 1986), Fodor (1987, 1990), Millikan (1984), and Papineau (1987, 1993).

property or relation. We will call this view *Hyper Representationalism*.[8]

3. <u>The Chomskian Challenge: *Ersatz Representationalism*</u>

Noam Chomsky has argued that the so-called 'representational' states invoked in accounts of our cognitive capacities are not correctly construed as about some represented objects or properties. Discussing computational vision theory he says,

> There is no meaningful question about the "content" of the internal representations of a person seeing a cube under the conditions of the experiments... or about the content of a frog's "representation of" a fly or of a moving dot in the standard experimental studies of frog vision. No notion like "content", or "representation of", figures within the theory, so there are no answers to be given as to their nature. The same is true when Marr writes that he is studying vision as "a mapping from one representation to another..." (Marr, 1982, p.31) – where "representation" is not to be understood relationally, as "representation of". (1995, 52-3)

The idea that "representation" should, in certain contexts, not be understood relationally, as in "representation of x", but rather as specifying a monadic property, as in "x-type representation", can be traced to Goodman 1968. So understood, the individuating condition of a given internal structure is not its relation to an 'intentional object', there being no such thing according to Chomsky, but rather its role in cognitive processing. Reference to what

[8] Dretske, a proponent of Hyper Representationalism, adds the requirement that content has causal powers, in some sense. He expects an adequate theory of representation to explain how content "gets its hands on the steering wheel." [Dretske, 1988]

looks to be an intentional object is simply a convenient way of type-identifying structures with the same role in computational processing.

The point applies as well to the study of the processes underlying linguistic capacities:

> The internalist study of language also speaks of 'representations' of various kinds, including phonetic and semantic representations at the 'interface' with other systems. Here too we need not ponder what is represented, seeking some objective construction from sounds or things. The representations are postulated mental entities, to be understood in the manner of a mental image of a rotating cube, whether the consequence of tachistoscopic presentations or of a real rotating cube or of stimulation of the retina in some other way, or imagined, for that matter. Accessed by performance systems, the internal representations of language enter into interpretation, thought, and action, but there is no reason to seek any other relation to the world... (Chomsky, 1995, 53)

Chomsky rejects the idea that intentional attribution – the positing of a domain of objects or properties to which internal structures stand in a *meaning* or *reference* relation – plays any explanatory role whatsoever in cognitive science. Characterizing a structure as 'representing an edge' or 'representing a noun phrase' is just loose talk, at best a convenient way of sorting structures into kinds determined by their role in processing. We shouldn't conclude that the structure is a representation *of* anything. Intentional construals of David Marr's 1982 theory of vision, such as Burge 1986 and many subsequent accounts, Chomsky claims, are simply a misreading, based on conflating the theory proper with its informal presentation. As Chomsky puts it, "The theory itself has no place for the [intentional] concepts that enter into the informal presentation, intended for general motivation." (1995, 55)

Chomsky's 'non-relational' representation is representation in a very *minimal* sense – representation without a domain of represent*ed* objects or properties. Recall the earlier depiction of an adder. The bottom span depicts the representational *vehicles*, the top span the represented *objects*. Chomsky's idea is to lop off the top span, leaving structures that don't represent anything. Recall our earlier definition of 'representation' (from Markman and Dietrich 2000): "mediating states of an intelligent system that carry information." In Chomsky's hands the notion is simply "mediating states of an intelligent system." But an unfortunate consequence of Chomsky's view is that too many things will count as representations: intelligent systems have *all sorts* of mediating states. Surely they are not all representations.

I will call Chomsky's account of representation without represent*ed* objects (or properties) *Ersatz Representationalism*. While I think the notion of 'non-relational' representation is problematic at best (and possibly incoherent), it is nonetheless worth trying to understand the motivation for Chomsky's view.

One of Chomsky's goals is to dispel talk of cognitive systems 'solving problems' (as in 'system x evolved to solve problem y'), and related talk of 'misperception,' 'misrepresentation,' and 'error.' Such intentional notions, he claims, have no significant place in scientific theorizing about the mind. Not only are these notions overtly *normative*, but they also reflect what Chomsky regards as our *parochial* interests. These interests can be addressed within what he calls our culture's 'ethnoscience', but they have no place in science itself.

We might wonder why Chomsky persists in calling the structures posited in scientific accounts of cognition 'representations'. One motivation might be to maintain the appearance of consistency with his earlier views.[9] But more importantly, I will argue, Chomsky needs to retain some notion of representation to preserve the idea that cognitive theories describe some cognitive *capacity* or *competence*, since these, the explananda of scientific cognitive theories, are described in intentional terms. But the *ersatz* notion he proposes does not allow him to do this. In what follows I will sketch the notion of representation that, I will argue, computational cognitive science both needs and actually uses. It bears little resemblance to the notion characterized by Hyper Representationalism. We will see that there is something right about Chomsky's claim that representationalist talk is "informal presentation, intended for general motivation" (1995, 55), even though, as construed by Chomsky himself, representationalist talk is unable to play this role.[10] In the next section I will discuss very briefly two computational accounts from different cognitive domains. These examples will enable me, in the section that follows, to sketch an alternative account of representation that adverts to two kinds of content.

4. Two examples

The first example is from Marr's well-known theory of early vision. Marr 1982 describes a component of early visual processing responsible for the initial filtering of the retinal image. The device takes as input light intensity values at points in the image and calculates the rate of change of intensity over the image. The implementation of this function is used in Marr

---

[9] He wrote a book in the 1970s called *Rules and Representations*. Though see Collins (2007) for the view that Chomsky has always been an anti-representationalist.

[10] Chomsky sometimes suggests that we should dispense with representationalist talk altogether:

> I do not know of any notion of 'representational content' that is clear enough to be invoked in accounts of how internal computational systems enter into the life of the organism. (2003, 274)

and Hildreth's (1980) model of edge detection to detect zero-crossings. (A zero-crossing is a point where the value of a function changes its sign. Zero-crossings correspond to sharp intensity changes in the image.) Although there are many ways to informally describe what the filter does, Marr is careful to point out that the theoretically important characterization, from a computational point of view, is the *mathematical* description: the device computes the Laplacean convolved with a Gaussian (1982, p. 337). The *canonical* description of the task executed by the device – the description that type-individuates it and distinguishes it from other computational mechanisms – is the *mathematical* description.

The second example, from a different cognitive domain, is Shadmehr and Wise's (2005) computational theory of motor control. Consider a simple task involving object manipulation. (See figure 1) A subject is seated at a table with eyes fixated ahead. The hand or *end effector* (ee) is located at Xee, and the target object (t) at Xt. The problem is simply how to move the hand to grasp the object. There are an infinite number of trajectories from the hand's starting location Xee to the target at Xt. But for most reaching and pointing movements, the hand moves along just one of these trajectories, typically, a straight path with a smooth velocity. Shadmehr and Wise describe one way the task might be accomplished.

The overall problem can be broken down into a number of sub-problems. The first problem is *how does the brain compute the location of the hand*? *Forward kinematics* involves computing the location of the hand (Xee) in visual coordinates from proprioceptive information from the arm, neck, and eye muscles, and information about the angles of the shoulder and elbow joints. Informally, this process coordinates the way the hand looks to the subject with the way it feels. The brain also has to compute the location of the target (Xt), using retinal information and information about eye and head orientation.

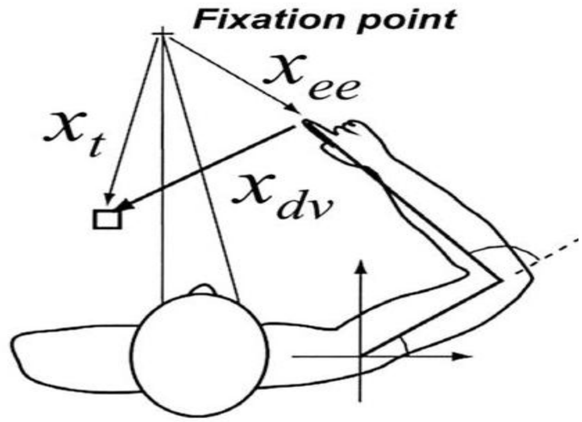


Figure 1

The second problem, computing a plan of movement, involves computing the *difference vector*, that is, the displacement of the hand from its current location to the target's location. But this 'high level' plan specifies a displacement of the hand in visual coordinates. The visually oriented plan has to be transformed into a specification of the joint rotations and muscle forces required to effect the displacement. So, the third problem, involving the computation of *inverse kinematics and dynamics*, is how the high level motor plan, corresponding to a difference vector, is transformed into joint angle changes and force commands. Reaching and pointing movements involve continuous monitoring of target and hand location, with the goal of reducing the difference vector to zero. There are a number of complicating factors. Incidental eye and head movements require continuous updating of the situation. Deceleration of the hand should be smooth, to avoid knocking over the target.

Summarizing, the account decomposes the overall task into three computations, and specifies the function computed in each in precise mathematical terms:

(1) $f(\theta) = X_{ee}$, *forward kinematics*, the computation of hand location, in eye-centered coordinates, from propriocentric information and information about joint angles;

(2) $X_t - X_{ee} = X_{dv}$, *the difference vector*, the difference between the target location and initial hand position in eye-centered coordinates; and

(3) $f(X_{dv}) = \Delta\theta$, *inverse kinematics*, the computation from the high-level movement plan, in eye-centered coordinates to a required change of joint angles.

The three computations that constitute the motor control mechanism are characterized by Shadmehr and Wise as analog processes and realized in neural networks in the posterior parietal cortex, the premotor cortex, and the primary motor cortex respectively. The details need not concern us here.

5. Some morals

I shall now draw some explicit morals from the examples:

1. The Shadmehr/Wise model of motor control explains our capacity to grasp objects in our immediate environment. Marr's account of the visual filter helps to explain our ability to see 'what is where' (as Marr puts it) in the nearby environment. Importantly, it is *assumed* from the outset that we are successful at these tasks; this success is the explanandum of the theory. The question for the theorist is *how* we do it. The significance of the assumption of success will emerge below.

2. In both examples – the visual mechanism and the motor control mechanism – the canonical description of the task executed by the device, the function(s) computed, is a mathematical description. This description characterizes the mechanism as a member of a well-understood class of mathematical devices. This is an essential feature of these accounts: they

allow us to bring to bear knowledge about how such functions can be executed. This mathematical characterization – which I will call a *function-theoretic* characterization – gives us a deeper understanding of the device; we *already* understand such mathematical functions as vector subtraction, Laplacian of Gaussian filters, integration[11], etc. Shadmehr and Wise's characterization of the motor-control mechanism allows us to see that a mariner who knew the distance and bearing from his home port to his present location and the distance and bearing from his home port to a buried treasure could perform the *same* computation – vector subtraction – to compute the course from his present location to the treasure.

A crucial feature of the function-theoretic characterization is that it is 'environment neutral': the task is characterized in terms that prescind from the environment in which the mechanism is normally deployed. The mechanism described by Marr would compute the Laplacean of the Gaussian even if it were to appear (*per mirabile*) in an environment where light behaves very differently than it does on earth, or as part of an envatted brain. It would compute this function whether it is part of a visual system or an auditory system, in other words, independently of the environment – even the *internal* environment – in which it is normally embedded. In fact, it is not implausible to suppose that each sensory modality has one of these same computational mechanisms, since it just computes a particular curve-smoothing function, a computation that may be put to a variety of different cognitive uses in different contexts. In some internal environments it would sub-serve vision; in a different internal environment, it might sub-serve audition. The function-theoretic description, then, provides a *domain-general* characterization of the device. Even a relatively simple device such as the Marrian filter, which computes only the Laplacean of a Gaussian, is, in this sense,

---

[11] See, for example, the account of oculomotor mechanisms in Seung 1996 and Seung et al 2000.

a *multi-purpose* tool.[12]

3. There is a kind of content that is essential in the two accounts, but it is *mathematical* content. This is one of the two kinds of content that I alluded to at the beginning. Inputs to the visual filter – the 'gray level array' – represent numerical values over a matrix. Outputs represent rate of change over the matrix. Inputs to the component of the Shadmehr/Wise mechanism that computes vector subtraction represent vectors and outputs represent their difference. More generally, the inputs of a computationally characterized mechanism represent the *arguments* and the outputs the *values* of the mathematical function that canonically specifies the task executed by the mechanism.

Let us pause for a moment to consider such mathematical contents in light of the view be naturalized, as Hyper Representationalism requires. What robust relation, specifiable in non-semantic and non-intentional terms, holds between the structures that make up the gray-level array and (just) the mathematical objects to which they are mapped in the interpretation? Indeed, it is hard to see what naturalization would amount to here. At very least, the naturalistic proposals currently on offer – information-theoretic accounts that depend on a causal relation, teleological accounts that advert to evolutionary function – are non-starters.[13] Whether or not there is such a content-determining relation, the success (and, I would argue, the legitimacy) of computational theorizing does not depend on it.

The focus on the function-theoretic characterization of a computational mechanism prompts the following question: How does computing the specified mathematical function

---

[12] This sense of *function-theoretic* characterization is not to be confused with various other notions in the literature, in particular, with Cummins' (1975) notion of *functional analysis*.
[13] Of course, it does *not* follow from the absence of a teleological relation grounding the ascription of mathematical content that the mechanism that computes the specified mathematical function does not thereby contribute to the fitness of the organism, i.e. that it is not an adaptation. The mechanism itself has a teleological explanation.

enable the mechanism to carry out its *cognitive* task – for the motor control mechanism, grasping an object in nearby space, for the Marrian filter, seeing 'what is where' in the near-by environment? As noted above, the function-theoretic description provides an environment-neutral, domain-general characterization of the mechanism. The theorist must explain how computing the value of the specified function, in the subject's normal environment, contributes to the exercise of the cognitive capacity that is the explanatory target of the theory. Only in *some* environments would computing the Laplacean of a Gaussian help the organism to see. In our environment this computation produces a smoothed output that facilitates the detection of sharp intensity gradients across the retina, which, when these intensity gradients co-occur at different scales, correspond to physically significant boundaries in the scene. One way to make this explanation perspicuous is to talk of inputs and outputs of the mechanism as *representing* light intensities and discontinuities of light intensity respectively; in other words, to attribute contents that are appropriate to the relevant cognitive domain, in this case, *vision*. At some point the theorist needs to show that the computational/mathematical account addresses the explanandum with which he began. And so theorists of vision will construe the posited mechanisms as representing properties of the light, e.g. light intensity values, changes in light intensity, and, further downstream, as representing changes in depth and surface orientation. Theorists of motor control will construe the mechanisms they posit as representing positions of objects in nearby space and changes in joint angles.

We will call the contents that are specific to the cognitive task being explained *cognitive contents*. This is the second kind of content I alluded to at the beginning. We will call the mapping that specifies these contents the *cognitive interpretation*. Cognitive contents, which are assigned for these explicative/elucidatory purposes, have the following properties:

1. *The explanatory context fixes the domain of cognitive interpretation*. The theorist may look for a distal causal antecedent of an internal structure's tokening, or a homomorphism between internal and distal elements, but the search is constrained primarily by the pre-theoretic explanandum, that is, by the cognitive capacity that the theory is developed to explain. A vision theorist assigns *visual* contents to explain the organism's capacity to see what is where in the scene, and so the theorist must look to properties that can structure the light in appropriate ways.

2. *No naturalistic relation is likely to pick out a single, determinate content*. Any number of relations may hold between the representing state or structure (the vehicle) and the object or property to which it is mapped in the cognitive interpretation. There may be a causal relation between states of the device and elements of the target domain, but there will be other candidates – distal and proximal – in the causal history of the process, equally good from a naturalistic perspective.[14] There will generally be a homomorphism between states of the system and the target domain, as between elements in a map and what they represent, but, famously, homomorphisms are cheap; there will be *many* such mappings. The important point is that no naturalistic relation determines the contents specified in computational models, as Hyper Representationalism requires.

3. *Use is crucial*. Even if some naturalistic relation were to (uniquely) hold between the posited structures and elements of the domain specified by the cognitive interpretation, the exist-

---

[14] To cite a well-known example, consider a frog that snaps out its tongue at any small dark thing moving in its visual field. Usually these are flies. But there are alternative candidates for the content of the structures constructed by a frog's visual system and appealed to in the explanation of the frog's feeding behavior: *fly*, *food*, *fly or BB*, *small dark moving thing*, *fly stage*, *undetached fly part*, etc. No purely naturalistic relation will privilege one of these candidates as '*the* object of perception.' And it is something of a fool's errand to try to decide just which is the single correct candidate. Various pragmatic considerations will motivate different content choices, as I explain below.

ence of this relation would not be sufficient to determine their cognitive contents. The structures have their cognitive contents only because they are *used* in certain ways by the device, ways that facilitate the cognitive task in question.[15] The fact that tokenings of the structure are regularly caused by some distal property tokening, and so can be said to 'track' that property, is part of the explanation of how a device that uses the posited structure in that way is able to accomplish its cognitive task (e.g. seeing what is where in the nearby environment), but the mere existence of the causal relation (or of an appropriate homomorphism) would not itself justify the content ascription in the absence of the appropriate use. It is also a matter of how the posited structures are used by processes that play a crucial role in the organism's capacity to see.

4. *In addition to the explanatory context – the cognitive capacity to be explained – other pragmatic considerations play a role in determining cognitive contents*. Given their role in explanation, candidates for content must be *salient* or *tractable*. The structure *EDGE*[16] in Marr's theory represents a change in depth, surface orientation, illumination, or reflectance, but if the causes of a structure's tokening are *too* disjunctive the theorist may decide to assign a *proximal* content to the structure (for example, zero-crossings represent discontinuities in the image), motivated in part by a desire to help us to keep track of what the device is doing at a given point in the process. In general, contents are assigned to internal structures constructed in the course of processing primarily as a way of helping *us* keep track of the flow of information in the system, or, more neutrally put, helping *us* keep track of changes in the system caused by both environmental events and internal processes, all the while with an eye on

---

[15] In so-called classical computational devices, this use is characterized as operations defined over explicit data structures.

[16] I use caps here to indicate the structure (the vehicle), independent of the content it is assigned in what I am calling the 'cognitive' interpretation.

the cognitive capacity (e.g. seeing what is where) that is the explanatory target of the theory. Contents play mainly an expository role. The choice of content will be responsive to such considerations as *ease of explanation*, and so may involve considerable idealization.

In Egan 2010 I develop a fictional example that makes explicit how 'ease of explanation' can play a significant role in the choice of a cognitive interpretation. The example is adapted from Segal (1989), though Segal, a proponent of narrow content, would not endorse the conclusions I draw from my use of the example. The example requires a bit of stage-setting.

Imagine a mechanism, Visua, which computes the depth of objects and surfaces in the immediate vicinity from information about the disparity of points in the retinal image. Visua is able to accomplish this, in part, because its states covary with changes in depth, or edges, in the immediate environment. The computational theory that characterizes Visua describes it in informal terms as representing *edges*.

Now imagine a physical duplicate of Visua – Twin Visua – in another environment, which we will call, for convenience, Twin Earth. Visua and Twin Visua are the same mechanism, from a computational point of view. They compute the same class of mathematical functions, using the same algorithms, with the same neural hardware. But Twin Earth is different enough from Earth – light behaves differently there – that Twin Visua's states do not co-vary with object boundaries there. However, they do co-vary with shadows, i.e. with changes in illumination, and tracking shadows on Twin Earth helps an organism to see, as Marr put it, what is where in its immediate surroundings. So Twin Visua, like Visua, is a *visual* mechanism.

Let's return to Earth. The theorist responsible for characterizing Visua has written a popular textbook on the computational theory of edge detection. An enthusiastic editor at MIT Press, always on the lookout for new markets, asks this theorist to produce a new edition of the text that could be marketed and sold on both Earth and Twin Earth. Visua and Twin Visua are computationally identical mechanisms – the function-theoretic description, the algorithms, and the physical implementation that characterizes Visua applies to Twin Visua as well. The theorist proposes a single cognitive interpretation that specifies what this mechanism represents in both worlds. Since the mechanism does not track shadows on Earth or edges on Twin Earth, neither *edge* nor *shadow* is a plausible candidate for the content. Rather, the proposed cognitive interpretation appropriate to both worlds takes the mechanism to represent some more general property – we will call it 'edgedows' – that subsumes both edges and shadows.

It is worth noting that the content *edgedow* is not a *narrow* content; it does not supervene on intrinsic properties of the subject and is not shared by all physical duplicates. It is a *distal* content. The new cognitive interpretation specifies what the mechanism represents on Earth and Twin Earth, but not what a physically (and computationally) indistinguishable mechanism might represent in some third, sufficiently different, environment Triplet Earth. (This follows by an iteration of the reasoning above.) While nonetheless wide or broad, *edgedow* is, in a sense, *narrower* than either *edge* or *shadow*. *Edgedow* prescinds from the environmental differences between Earth and Twin Earth. The explanatory interests served by the new interpretation are less local, less *parochial*, than those served by the original interpretation, which was designed to address questions posed in vocabulary appropriate to Earth. Whereas the original cognitive interpretation enabled the theory to address such pretheoretic questions as "how is the organism able to recover what is where in the scene" by positing representations of edges, the new interpretation provides the basis for answering this question by positing representations of the more general distal property *edgedow*, provided, of course, that this new interpretation supplies auxiliary assumptions about how *edgedow* is related to the locally instantiated properties edges (on Earth) and shadows (on Twin Earth). It is also worth noting that *edgedow* is no less kosher, from a naturalistic perspective, than *edge*.

As it happened, the editor was surprised when sales on Earth of the new interplanetary textbook fell off rather sharply from the first edition, designed solely for the local market. Besides introducing a new vocabulary containing such unfamiliar predicates as "edgedow", the new edition required cumbersome appendices appropriate to each world, explaining how to recover answers to questions about the organism's capacities in its local environment, questions that motivated the search for an explanatory theory in the first place. Readers complained that the new edition was much less "user-friendly".

The editor was therefore dissuaded from her original idea of commissioning an intergalactic version of the text, which would provide a genuinely narrow cognitive interpretation that would specify what Visua would represent in *any* environment.[17] She came to realize that a distal interpretation of a computationally characterized process is primarily a *gloss* that allows a theory to address local explanatory interests. Any gloss that shows that the theory is doing its job will be couched in a vocabulary that is perspicuous for the local audience with these interests. An important moral here is that a truly intergalactic computational cognitive science would not be *representational* in the following sense: it is not likely to assign any-

[17] Instead the editor commissioned an environment-specific cognitive interpretation for each world, to accompany the environment-neutral account of the mechanism provided by the computational theory.

thing that looks remotely like ordinary representational content.

Returning to our characterization of *cognitive content*:

5. *The assignment of cognitive content allows for misrepresentation, but only relative to a particular cognitive task or capacity*. In low light, a shadow may be mistaken for an object boundary (edge). In an Ames room (say, at Disney World), where the light is systematically distorted, the subject may misjudge the character of the local space. In such cases, the cognitive interpretation, which assigns visual contents, will specify a content – say, *edge* – which on this occasion is tokened in response to a shadow or some other distal feature. The mechanism *misrepresents* a shadow as an edge. All the while this mechanism computes the same mathematical function it always computes, but in an 'abnormal' situation (low light, distorted light, etc.) computing this mathematical function may not be sufficient for executing the cognitive capacity. Misrepresentation is something we attribute to the device when, in the course of doing its usual mathematical task (given by the function-theoretic description), it fails to accomplish the cognitive task specified by the pre-theoretic explanandum.[18]

6. *The structures posited by the computational theory, what we are calling the 'representational vehicles', do not have their cognitive contents essentially*. If the mechanism characterized in mathematical terms by the theory were embedded differently in the organism, perhaps allowing it to sub-serve a different cognitive capacity, then the structures would be assigned *different* cognitive contents. If the subject's normal environment were different (as, for example, in an Ames room), so that the use of these structures by the device in this environ-

[18] It is also possible for *mathematical* content to be non-veridical. If the device malfunctions, it may misrepresent the output of the specified function. If the function is defined on an infinite domain (as is, for example, the addition function) then the mathematical interpretation will involve some idealization. In general, the gap between competence (whose characterization will often involve idealization) and performance allows for the possibility of misrepresentation.

ment did not facilitate the execution of the specified cognitive task, then the structures might be assigned no cognitive contents at all. And the various pragmatic considerations cited above might motivate the assignment of different cognitive contents to the structures.[19]

7. *Cognitive contents are not part of the computational theory proper – they are part of the 'intentional gloss'.* What we might call the *computational theory proper* comprises the characterization of representational *vehicles*, the processes defined over them (which specify relevant aspects of representations' *use*), the specification of the mathematical function(s) computed by the device (what I am calling the 'function-theoretic' description), and the algorithms involved in the computation of this function. These components of a computational account provide an environment-independent characterization of the device. They therefore have considerable counterfactual power: they provide the basis for predicting and explaining the behavior of the device in *any* environment, including environments where the device would fail miserably at computing the cognitive function that it computes in the local ("normal") environment. But since the theory must explain the organism's manifest success at computing the cognitive function in its normal environment (e.g. seeing what is where, grasping objects in view), it must also advert to general facts about that environment that explain why computing the mathematical function specified by the theory, in context, suffices for the exercise of the cognitive capacity. Thus the 'theory proper' will also include such environment-specific facts as that a co-incidence of sharp intensity gradients at different scales is likely to be physically significant, corresponding to object boundaries or reflectance or illumination changes in the world.[20] Cognitive contents, on the other hand, are *not* part of

[19] These possible scenarios would not affect ascription of mathematical content.

[20] General environmental facts that Marr called *physical constraints* – such as that objects are rigid in translation (Ullman's (1979) *rigidity constraint*) or that disparity varies smoothly al-

the essential characterization of the device, and are not fruitfully regarded as part of the computational theory proper. They are ascribed to facilitate the explanation of the cognitive capacity in question and are sensitive to a host of pragmatic considerations, as explained above. Hence, they form what I call an *intentional gloss*, a gloss that shows, in a perspicuous way, how the computational/mathematical theory manages to explain the intentionally-described explanandum with which we began and which it is the job of the theory to explain.

To summarize: a computational account of a cognitive capacity can be partitioned into two parts: (1) the *computational theory proper*, which includes an environment-independent, function-theoretic characterization of the mechanism as well as general facts about the normal environment that are, strictly speaking, sufficient to explain the contribution of the mechanism to the organism's success in that environment; and (2) an *intentional gloss* that facilitates the foregoing explanation by characterizing the mechanism as *representing* elements of the target domain. The latter serves a heuristic purpose and is subject to a host of pragmatic considerations.

I have explained how this view differs from the view I have called Hyper Representationalism. Computationally characterized mental states, on the account I have sketched, do not have their representational contents (what I am calling their *cognitive* contents) essentially, and do not stand in some robust, naturalistic representation relation to what they are about. They do, however, satisfy the third HR constraint – they can *misrepresent*, but only relative to 'success criteria' provided by the cognitive capacity to be explained, that is, by the pre-theoretic explanandum. In the next section I return, briefly, to Chomsky's challenge. We will see that while his account is mistaken in some respects, he is nonetheless on to

---

most everywhere, since matter is cohesive (the *continuity constraint*) – are also part of the computational theory proper.

something important.

6. *Ersatz* Representationalism re-visited

Chomsky is wrong to claim that the internal structures posited in computational theories are not construed, in the theory, as representations of anything. They are construed in the theory as representations of mathematical objects, and they have their mathematical contents *essentially*. But this is just to say that a computational characterization is a *function-theoretic* characterization, in the sense I have explained. To characterize a mechanism as computing a function (in the mathematical sense) *just is* to construe its inputs and outputs as representing (respectively) the arguments and values of the function.

As Chomsky points out, to say that a mechanism 'solves a problem' or 'makes a mistake' is to give a *normative* characterization of its behavior. He complains that describing a device in such terms is to impose our own parochial interests and expectations on it. He thinks that such characterizations have no place in scientific inquiry. But there is an important distinction in empirical science between the theory proper and the apparatus for motivating the theory, showing how it explains the phenomena in its explanatory domain, and so on. For computational theories of cognitive capacities, as I have argued above, this apparatus includes an *intentional gloss*, and representational contents, as normally understood, are part of that gloss. Chomsky's argument appeals to a distinction between the theory proper and its 'informal presentation', but he disparages the latter as reflecting our parochial interests. He recognizes that the intentional characterization is part of a gloss, but fails to understand what the gloss is supposed to do.[21]

I mentioned earlier that Chomsky needs to retain some notion of 'representation' to preserve the idea that the theory describes a *cognitive capacity* or *competence*. The characterization of a phenomenon as a *capacity* or *competence* is itself normative, given pre-theoretically, prior to the theory's characterization of the mechanism underlying the competence. The normative elements are there from the beginning. The mechanisms characterized by cognitive theories must bear some fairly transparent relation to these pre-theoretically described capacities. In characterizing these mechanisms the computational theorist cannot let theory construction float free of the explanatory target – cognitive capacities, or, as it is often put, *solving problems* (e.g. figuring out what is where in the environment, understanding a stream of speech, grasping an object). The assignment of appropriate representational contents to the structures posited by the theory – taking them to be representations of edges or joint angles or noun phrases – secures the connection between the mechanism described in the theory in mathematical (and physical) terms, and the cognitive capacity (competence) that is the explanatory target.

Chomsky is right that such notions as *misrepresentation*, *error*, and *mistake* are not part of the computational theory proper. But it doesn't follow that these notions are dispensible. Representational contents – that is, what I have been calling *cognitive* contents – constitute a starting point for cognitive theory – specifying an intentional explanandum – and later a role, as part of what I am calling the *intentional gloss*, in justifying the theory, that is, in demonstrating that the theory has explained the capacity in question. Maybe it *is* rather paro-

[21] Hyper Representationalists (e.g. Fodor) do not recognize a distinction. They assume that there is just the theory proper, and that representational contents are part of it, giving an essential characterization of computational mechanisms and processes.

chial of us to want to see the processes that are described in non-intentional terms in the theory as constituting the exercise of a *capacity* or *competence*. Nonetheless, *pace* Chomsky, there is no reason why science should not aim to explain the features of our experience that interest us, even if it tells us that these features do not go very deep.[22]

7. Computational models and *intrinsic intentionality*

Intentional mental states are often said to differ from other things that have meaning – utterances, marks on the page or on the computer screen – inasmuch as they have their meanings not *derivatively*, i.e. not based on conventions governing their use, but rather *intrinsically* (or *originally*).[23] Much of the interest among philosophers of mind in computational cognitive science comes from its promise to provide a purely mechanical, or at least naturalistic (in some sense of the word), account of the mind. In so doing, metaphysicians of mind hope, it will explain intrinsic intentionality. In this penultimate section I will make some brief remarks on this prospect.

I have been at pains to argue that the states and structures characterized by computational theories do *not* have the earmarks of intrinsic intentionality. States with intrinsic intentionality have determinate content. But computationally characterized states have determinate content only by appeal to various pragmatic considerations such as ease of explanation and connections to our commonsense practices. (Recall the moral of the editor story, where

---

[22] It is not only the cognitive sciences that are grounded in our desire to understand ourselves and our place in the universe – the biological sciences are grounded in such interests as well. From the detached perspective of fundamental physics, the difference between life and non-living matter is no less arbitrary than the difference between a rational process and a mistake.
[23] Searle 1980, 1993.

these considerations constrain the choice among equally naturalistic alternatives.) We don't find the grounds of determinacy in the computational theory itself.

States with intrinsic intentionality are said to have their content essentially. But, at most, computationally characterized states and structures have their mathematical content essentially. We might be tempted to conclude that what they *really* represent are not distal objects and properties, but mathematical objects. But I don't think this is the right way to think of the matter. As I argue above, the canonical description of a mechanism given by a computational theory is function-theoretic, and to characterize a mechanism as computing a function (in the mathematical sense) *just is* to construe its inputs and outputs as representing (respectively) the arguments and values of the function. But even this 'essential' mathematical content has a pragmatic rationale – it reflects the commitment of computational practice to providing a (cognitive) domain-general characterization from a 'toolbox' of mathematical functions that are already well understood. The fact that the posited structures are assigned mathematical content in the theory doesn't preclude them having *different* contents in the gloss. Multiple content ascriptions serve different explanatory purposes. The idea that computational cognitive science is looking for, or fixing on, *the* content of mental states finds no support from actual theorizing.

Given that computational cognitive science aims to provide a *foundation* for the study of cognition it should not be surprising that we don't find full-blooded intrinsic intentionality in the theories themselves. Intrinsic intentionality is among the pretheoretic explananda of scientific theories of cognition. We should not look for intrinsic intentionality *in* computational models, any more than we should look for intrinsic *representations* – structures that have their representational roles independently of how they are used in specific contexts.

Computational cognitive science has reductive ambitions – it aims to *explain* the representational capacities of minds, without simply assuming these representational capacities. It would therefore be an explanatory *failure* if computational models posited states and structures that have all the marks of intrinsic intentionality. The ultimate goal is to explain how meaning and intentionality fit into the natural world. Explanatory progress is not made by positing more of the same mysterious phenomenon.

This is not to suggest that computational cognitive science has succeeded in reducing content and intentionality. Computational theories appeal to (unreduced) mathematical content. But we can see that a well-confirmed computational theory that included an account of how a mechanism is realized in neural structures would make some progress toward a reductive explanation of intentionality in the cognitive domain in question. What we normally think of as representational contents – contents defined on distal objects and properties appropriate to the cognitive domain (what I have called 'cognitive' contents) – are not in the theory; they are in the explanatory gloss that accompanies the theory, where they are *used* to show that the theory addresses the phenomena for which we sought an explanation. The gloss allows us to see ourselves as solving problems, exercising rational capacities, occasionally making mistakes, and so on. It characterizes the computational process in ways congruent with our commonsense understanding of ourselves, ways that the theory itself eschews.[24]

8. Postscript: can phenomenal experience save Hyper Representationalism?

[24] If we ever succeed in solving the so-called 'hard problem' of consciousness – providing a reductive explanation of phenomenal experience – we will undoubtedly need a *phenomenal gloss* that makes use of phenomenal concepts – concepts eschewed by the theory itself – to show that the theory addresses the phenomenon in question.

It has recently been suggested that rather than looking to external relations between mental states and distal objects or properties to ground determinate content, as Hyper Representationalists like Fodor propose, we should look *inside*, to the subject's phenomenal experience.[25] Indeed, Horgan and Graham (2012) claim that phenomenally-based intentionality is the source of all determinacy of thought content, even for the deeply unconscious, sub-doxastic, information-carrying states posited by computational theories of cognition. If they are right, then these states are *essentially* intentional after all, and the central plank of Hyper Representationalism is preserved. Here is Horgan and Graham's proposal:

> … we are fairly optimistic about the following hypothesis: given a specific cognitive-scientific account of the cognitive architecture of competent human cognizers, there will be a *unique* content-assignment C of intentional contents to internal states of humans that meets the following two constraints: (i) C assigns to phenomenally conscious states their determinate, inherent, phenomenal-intentional content, and (ii) C assigns contents to all other internal states in such a way that C exhibits an acceptably high degree of overall internal rational coherence (both synchronically and diachronically). The key idea here is that the phenomenally fixed contents of the phenomenally conscious mental states provide a sufficiently constraining network of "anchor points," for an overall assignment C of intentional contents to actual and potential internal states of human creatures, that only one such assignment can simultaneously honor all these anchor points and also render rationally appropriate, both for total synchronic states and for diachronic state-transitions, all the assigned contents. (p.341)

[25] For a very useful discussion of the attempt to ground intentionality in phenomenal character see Kriegel (2013).

One motivation for Horgan and Graham's view is their desire to "embrace a view regarding the foundations of cognitive science that can smoothly accommodate any kinds of deeply-unconscious mental states whose existence might get posited by otherwise well-warranted cognitive-scientific theories" (p.341). They mention explicitly Marr's computational theory of early vision. I shall leave aside Horgan and Graham's claim [re (i)] that all phenomenally conscious states have determinate content in virtue of their intrinsic phenomenal character and comment only on their claim that phenomenal consciousness will fix determinate content for the states and structures posited in cognitive scientific theories.

Horgan and Graham's view of how the content of non-conscious states posited by cognitive science is fixed is clearly at odds with my account of computational theorizing. I have argued that computational accounts posit two kinds of content – mathematical contents in the theory and cognitive contents in the explanatory gloss. The ascription of mathematical content is not constrained by phenomenal considerations. The search for a function-theoretic characterization of a cognitive mechanism is motivated by a desire to subsume the mechanism under a wider class of well understood, abstract devices (e.g. Laplacean/Gaussian filters, integrators, etc.). By its very nature, a function-theoretic characterization subsumes both biological and artifactual computers. The latter (we may presume) have no phenomenality, but they would be assigned the same mathematical content by the theory.

Turning to cognitive contents, recall that the process characterized in Shadmehr and Wise's theory of motor control as *forward kinematics* is described informally as 'coordinating the way the hand looks to the subject with the way it feels.' Phenomenal experience can play a rich role in specifying the pre-theoretic explananda of cognitive theories, and so a rich role in the explanatory *gloss* that accompanies a computational theory. Certainly, phenomenality can play a larger role in the specification of cognitive contents than allowed by theories that attempt to ground determinate content in external 'naturalistic' relations alone (that is, by theories favored by Hyper Representationalists). However, there is no reason to think that phenomenal experience will provide sufficient "anchor points" to determine unique contents for the states and structures posited in computational theories, in the absence of explanatory and other pragmatic considerations that I describe above (which will include various relations to distal properties in the subject's normal environment).

Let's put my account of content aside. Horgan and Graham's view that phenomenal consciousness will fix determinate content for the non-conscious states posited in cognitive theories finds no support in the actual practice of cognitive researchers. Theorists of cognition typically look to the organism's behavior and to the environment in which the behavior is normally deployed to characterize what the posited states are doing. They look to characteristic patterns of error. Interpreters of Marr's theory, who disagree about how the computational specification is to be understood and about what type of content is ascribed in the theory, agree on that much.[26] In what is widely regarded as a classic of cognitive theorizing, Gallistel (1990) describes the foraging behavior of the Tunisian desert ant by positing in the ant a Cartesian representation of its current position. He arrives at this description by examining the ant's behavior in its desert environment, taking into account patterns of success and failure. Even if ants have phenomenal states, phenomenal considerations play no role in the theory. While Gallistel's theory is about the Tunisian desert ant, his methodology for the study of cognition is completely general. Horgan and Graham claim only that phenomenal consciousness will provide a unique assignment of content for all actual and potential internal

[26] See, for example, Burge (1986), Segal (1989), Davies (1991), Shapiro (1993), Egan (1995, 1999), and Shagrir (2001, 2010).

states of *human* subjects, but the scientific study of animal and human cognition is continuous. Cognitive mechanisms are likely to be *adaptations*; many of them are likely to be found, perhaps only with relatively minor variation, in non-human animals. (This is especially likely for the mechanisms responsible for our perceptual and motor capacities.) There is no reason to take seriously the idea that the content of non-conscious internal states posited by our most promising theories of cognition is fixed, or even substantially constrained, by phenomenal considerations.[27] [28]

---

[27] Horgan and Graham express a desire to "leave behind what Searle… tantalizing calls 'the persistent objectivizing tendency of philosophy and science since the seventeenth century' (Searle 1987, 145)" (p.18), but the fact is that the cognitive sciences in general, and computational cognitive science in particular, are firmly in the tradition of post-Galilean science.
[28] Thanks to Todd Ganson, Mohan Matthen, Robert Matthews, and participants at the Oberlin Philosophy Conference, May 2012 for comments on earlier versions of this paper.

References

Burge, T. (1986), “Individualism and Psychology,” *The Philosophical Review*, 95: 3-45.

Chomsky, N. (1995), “Language and Nature,” *Mind*, 104: 1-61.

Chomsky, N. (2003), “Reply to Egan,” In L.Antony and N. Hornstein (eds.), *Chomsky and his Critics*, Oxford: Blackwell, 89-104.

Collins, J. (2007), Meta-scientific eliminativism: a reconsideration of Chomksy’s review of Skinner’s *Verbal Behavior*. *British Journal for the Philosophy of Science*, 58: 625-58.

Cummins, R. (1975), “Functional Analysis,” *The Journal of Philosophy*, 72: 741-64.

Cummins, R. (1989), *Meaning and Mental Representation*, Cambridge, MA: MIT Press.

Davies, M. (1991), “Individualism and Perceptual Content,” *Mind*, 100: 461-84.

Dretske, F. (1981), *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

Dretske, F. (1986), "Misrepresentation," in *Belief: Form, Content, and Function*, R. Bogdan, ed., Oxford: OUP.

Dretske, F. (1988), *Explaining Behavior*, Cambridge, MA: MIT Press.

Egan, F. (1995), “Computation and Content,” *The Philosophical Review*, 104: 181-203.

Egan, F. (1999), “In Defense of Narrow Mindedness,” *Mind and Language*, 14: 177-94.

Egan, F. (2010), “Computational Models: A Modest Role for Content,” *Studies in History and Philosophy of Science* (special issue on *Computation and Cognitive Science*), 2010, 41: 253-259.

Fodor, J.A. (1987), *Psychosemantics*, Cambridge, MA: MIT Press.

Fodor, J.A. (1990), “A Theory of Content II: The Theory,” in *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, 89-136.

Fodor, J. (2008), *LOT2: The Language of Thought Revisited*, Oxford University Press.

Gallistel, R.C. (1990), *The Organization of Learning*, Cambridge, MA: MIT Press.

Goodman, N. (1968), *Languages of Thought*, Cambridge, MA: Harvard University Press.

Horgan, T. and Graham, G. (2012), “Phenomenal Intentionality and Content Determinacy,” in *Prospects for Meaning*, R. Schantz (ed.), Boston: De Gruyter, 321-344.

Kriegel, U. (2013), “The Phenomonal Intentionality Research Program,” in *Phenomenal Intentionality*, U. Kriegel (ed.), New York: Oxford University Press.

Markman, A, and Dietrich, E. (2001), “Extending the Classical View of Representation,” *Trends in Cognitive Science*, 470-475.

Marr, D. (1982), *Vision*, New York: Freeman.

Marr, D. and Hildreth, E. (1980), “Theory of Edge Detection,” *Proceedings of the Royal Society*, 187-217.

Millikan, R. (1984), *Language, Thought, and Other Biological Categories*, Cambridge, MA: MIT Press.

Papineau, D. (1987), *Reality and Representation*, Oxford: OUP.

Papineau, D. (1993), *Philosophical Naturalism*, Oxford: Blackwell.

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

Searle, J. (1980), ‘Minds, Brains and Programs’, *Behavioral and Brain Sciences*, 3: 417–57.

Searle, J. (1987), “Indeterminacy, Empiricism, and the First Person,” *The Journal of Philosophy* 84, 1123-146.

Searle, J. (1993), *The Rediscovery of The Mind*, Cambridge, MA: MIT Press.

Segal, G. (1989), “Seeing what is not there,” *The Philosophical Review* 98: 189-214.

Shadmehr, R. and Wise, S. (2005), *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA: MIT Press.

Shagrir, O. (2001), "Content, Computation and Externalism," *Mind*, 110: 369-400.

Shagrir, O. (2010), "Marr on Computational-Level Theories," *Philosophy of Science*, 77: 477-500.

Shapiro, L. (1993), “Content, Kinds, and Individualism in Marr’s Theory of Vision,” *The Philosophical Review*, 102: 489-513.

Sueng, H. S. (1996), “How the Brain Keeps the Eyes Still,” *Proceedings of the National Academy of Science (USA)*, 13339-13344.

Sueng, H. S., Lee, D., Reis, B., and Tank, D. (2000), “Stability of the Memory of Eye Position in a Recurrent Network of Conductance-Based Model Neurons,” *Neuron*, 26, 259-271.

Ullman, S. (1979), *The Interpretation of Visual Motion*, Cambridge, MA: MIT Press.