18

THE NATURE AND FUNCTION OF CONTENT IN COMPUTATIONAL MODELS

Frances Egan

Introduction

Much of computational cognitive science construes human cognitive capacities as representational capacities, or as involving representation in some way. Computational theories of vision, for example, typically posit structures that represent edges in the distal scene. Neurons are often said to represent elements of their receptive fields. Despite the ubiquity of representational talk in computational theorizing there is surprisingly little consensus about how such claims are to be understood. The point of this chapter is to sketch an account of the nature and function of representation in computational cognitive models.

A commitment to representation presupposes a distinction between representational *vehicle* and representational *content*. The vehicle is a physically realized state or structure that carries or bears content. Insofar as a representation is causally involved in a cognitive process, it is in virtue of the representational vehicle. A state or structure has content just in case it represents things to be a certain way; it has a 'satisfaction condition' – the condition under which it represents accurately.

The representational vehicles in so-called 'classical' computational systems are *symbols*, physical structures characterized by a combinatorial syntax, over which computational processes are defined. Symbols are tailor-made for semantic interpretation, for 'hanging' contents on, so to speak. But not all computational systems are symbol-manipulating systems. For example, connectionist models explain cognitive phenomena as the propagation of activation among units in highly connected networks; dynamical models characterize cognitive processes by a set of differential equations describing the behavior of the system over time. The systems so described do not operate on symbols in any obvious sense. There is a good deal of controversy about whether these systems are genuinely representational. For the most part, the dispute concerns whether such systems have representational vehicles, that is, states or structures causally involved in cognitive processes that are plausibly construed as candidates for semantic interpretation. In this chapter we will put this issue aside, abstracting away from questions about the bearers of content, and focus on the nature and function of representational content itself.

Adequacy conditions on an account of content for computational systems

We can identify several widely accepted constraints on an account of content for computational neuroscience:

1 The account should provide the basis for the attribution of *determinate* contents to computational states or structures.
2 The account should allow for the possibility that the posited states can *mis*represent.

The idea is that genuinely representational states represent *robustly*, in the way that paradigmatic mental states such as beliefs represent; and they should allow for the possibility of *getting it wrong*.

There is a constitutive connection between constraints (1) and (2). If the theory cannot underwrite the attribution of determinate satisfaction conditions to a mental state (type), then it cannot support the claim that some possible tokenings of the state occur when the conditions are not satisfied, and hence would misrepresent. For example, suppose that we want to capture the idea that a frog's tongue snapping at a BB in the laboratory constitutes a misrepresentation. This requires excluding BB-caused tokenings from the content-determining conditions of the internal state. Partitioning possible tokenings of the state into veridical instances on the one hand and misrepresentations on the other requires an antecedent specification of the state's content.

3 The account should be *naturalistic*.

Typically, this constraint is construed as requiring a specification, in non-semantic and non-intentional terms, of (at least) a sufficient condition for a state or structure to have a particular content. Such a specification would guarantee that the theory makes no illicit appeal to the very phenomenon – meaning – that it is supposed to explain. This idea motivates so-called *tracking* theories, discussed below. But we will see that this is not the only way to interpret the naturalistic constraint. More generally, the constraint is motivated by the conviction that intentionality is not *fundamental*:[1]

> It's hard to see … how one can be a realist about intentionality without also being, to some extent or other, a reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe supervenience on) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be something else.
>
> *(Fodor, 1987, p. 97)*

> There are no "ultimately semantic" facts or properties, i.e. no semantic facts or properties over and above the facts and properties of physics, chemistry, biology, neurophysiology, and those parts of psychology, sociology, and anthropology that can be expressed independently of semantic concepts.
>
> *(Field, 1975, p. 386)*

Philosophers of mind of a materialistic bent have traditionally been interested in computationalism in part because it seeks to characterize mental processes as *mechanical* processes – processes guaranteed to be physically realizable, whose specification invokes no mysterious mental substance, properties, or events. Its success would pave the way for a naturalistic reduction of the mind, and so of intentionality. Or so it has been hoped..

Finally,

4 The account should conform to actual practice in computational cognitive science.

Proposals for computational content

Let us turn now to the central question: how do states/structures posited in computational models get their meaning? I will discuss some popular proposals and indicate outstanding problems with each before sketching what I take to be the correct view. My discussion of the popular candidates will necessarily be very brief.

Tracking theories

Most theories of content explicate intentionality in terms of a privileged relation between the tokening of an internal state and what the state represents. Thus the state is said to 'track' (in some specified sense) the external condition that serves as its satisfaction condition. Tracking theories are explicitly naturalistic – both the relation and the relata should be specified in non-intentional and non-semantic terms – but they differ in their accounts of the representation relation.

Information-theoretic theories

Very roughly, according to information-theoretic accounts, an internal state S means *cat* if S is caused by the presence of a cat, and certain further conditions obtain.[2] Further conditions are required to allow for the possibility of misrepresentation, that is, for the possibility of some S-tokenings *not* caused by cats but, say, by large rats on a dark night. A notable problem for information-theoretic theories is the consequence that everything in the causal chain from the presence of a cat in the distal environment to the internal tokening of S, including cat-like patterns in the retinal image, may appear to satisfy the condition, and so falls into S's extension. Thus, information-theoretic theories typically founder on constraint (1), failing to underwrite determinate contents for mental states. The outstanding problem for such theories is to provide for determinacy without illicit appeal to intentional or semantic notions. Yet further conditions may sufficiently constrain content but if the proposed meaning-determining relation becomes too baroque it will fail to be explanatory, leaving us wondering *why* it determines content.

Teleological theories

According to teleological theories, internal state S means *cat* if and only if S has the natural function of indicating cats. The view was first developed and defended in Millikan (1984), and there are now many interesting variations on the central idea.[3] Teleosemanticists have been notoriously unable to agree on the natural function of states of even the simplest organisms.[4] Let's focus on a widely discussed case. Does the inner state responsible for engaging a frog's tongue-lashing behavior have the function of indicating (and hence representing) *fly*, *frog food*, or *small dark moving thing*? Teleosemanticists, at various times, have proposed all three. Suppose we settle on *fly*. Wouldn't a *fly stage* detector or an *undetached fly part* detector serve the purpose of getting nutrients into the frog's stomach equally well?[5] The problem is that indeterminate functions cannot ground determinate contents. Each of various function-candidates specifies a different satisfaction condition; unless a compelling case can be made for one function-candidate

over the others, teleosemantics runs afoul of constraint (1). Moreover, the argument should not appeal to intentional or normative considerations (such as what makes for a good explanation), on pain of violating the naturalistic constraint.

Structural similarity theories

A third type of tracking theory appeals to the type of relation that holds between a map and the domain it represents, that is, structural similarity or isomorphism. Cummins (1989), (Ramsey 2007), and Shagrir (2012) have proposed variations on this idea. Of course, since similarity is a symmetric relation but the representation relation is not, any account that attempts to ground representational content in similarity will need supplementation by appeal to something like *use*. Moreover, as the saying goes, "isomorphisms are cheap". A given set of internal states or structures is likely to be structurally similar to any number of external conditions. The question is whether structural similarity can be sufficiently constrained to underwrite determinate contents while still respecting the naturalistic constraint.

The upshot of this short discussion is that tracking theories face formidable problems, but it would certainly be premature to write them off. One might simply conclude that more work needs to be done. It is worth noting, however, that despite the fact that there is no widely accepted naturalistic foundation for representational content, computational theorists persist in employing representational language in articulating their models. For example, vision theorists talk of structures posited in the course of visual processing *representing* edges in the scene. Neuroscientists talk of cells in the hippocampus ('place cells') *representing* locations in the local environment. The apparent mismatch between the theories of content developed by philosophers pursuing the naturalistic project and the actual practice of computational theorists in ascribing content cries out for explanation; it motivates a different sort of account. Before sketching an account that better fits the practice I shall consider very briefly a couple of other proposals.

Phenomenal intentionality

It has recently been suggested by proponents of the *phenomenal intentionality research program* (PIRP)[6] that rather than looking to external relations between states of the subject and distal objects or properties to ground determinate content, as tracking theorists propose, we should look inside to the subject's phenomenal experience. 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-personal states posited by computational theories of cognition. Intriguing though the suggestion is, it has a number of problems. PIRP theorists reject the naturalistic constraint, as it is normally understood, but there are more serious worries. In the first place, the view finds no support in the actual practice of computational theorists, who typically look to an organism's behavior and to the environment in which the behavior is normally deployed when they assign representational content to computational states. They look to characteristic patterns of error. On a smaller scale, neuroscientists look to features of a neuron's receptive field. They do not look to the way things seem to the subject; though, of course, their theories often have implications for the subject's phenomenal experience. So the suggestion fails to comply with constraint (4), the requirement that an account of representational content should conform to actual practice in computational cognitive science. Second, and more importantly, whatever the current state of practice, the proposal is at odds with a fundamental commitment of computationalism, viz. the idea that thought is, at bottom, a *mechanical* process. This commitment underwrites the promise of artificial intelligence. Maybe a mechanical

account of phenomenal consciousness will eventually be forthcoming, explaining not only our own phenomenal experience but also paving the way for the creation of machine consciousness. But if so, the phenomenal intentionality program gets the grounding relation backwards. It will be computation that fixes the determinate content of phenomenal experience, not the other way around.

Content eliminativism

A natural reaction to the problem of grounding content might be to reject content altogether, as Noam Chomsky does.[7] According to Chomsky, characterizing an internal 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. 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, p. 55). In a later work he goes on to say:

> 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. And to the extent to which I can grasp what is intended, it seems to me very questionable that it points to a profitable path to pursue.
>
> *(Chomsky, 2003, p. 274)*

Chomsky is on to something important here. Structures posited by computational theories *are* sorted into kinds by their role in processing. And though they are often characterized by their contents, to take representational talk too seriously *is* to conflate the theory with its informal presentation. But content eliminativism doesn't follow, because Chomsky is wrong to conclude that content plays no explanatory role in computational cognitive models. These claims will be defended in the next section.

A deflationary account of content

The most popular accounts of content for computational theorizing – tracking theories – share two central commitments:

1 Mental representations have their contents *essentially*: if a particular internal structure had a different content it would be a different (type of) representation. In other words, computational theories individuate the states and structures they posit partly in terms of their content.
2 Content is determined by a privileged naturalistic relation holding between a state/structure and the object or property it is about.

In this section I will sketch an alternative picture of the nature and function of representational content in computational theorizing[8] – what I call a *deflationary* account of content – characterized by the rejection of the above two claims.

I begin by calling attention to two central features of computational theorizing:

1 Computational theories of cognitive capacities provide what I call a *function-theoretic* characterization of the capacity.
2 A computational theory (including the function-theoretic characterization, and the specification of algorithms, structures and processes) is accompanied by what I call an *intentional gloss*.

These two features, to be spelled out below, determine two kinds of content that play distinctive roles in computational theorizing.

Mathematical content

Marr's (1982) theory of early vision purports to explain edge detection, in part, by positing the computation of the Laplacean of a Gaussian of the retinal array. The mechanism takes as input intensity values at points in the image and calculates the rate of intensity change over the image. In other words, it computes a particular smoothing function. Marr's theory is typical of perceptual theories in this respect: perceptual systems compute smoothing functions to eliminate noise. Shadmehr and Wise's (2005) computational account of motor control explains how a subject is able to grasp an object in view by computing the displacement of the hand from its current location to the target location, i.e. by computing vector subtraction. Seung et al. (1996; 1998; 2000) hypothesize that the brain keeps track of eye movements across saccades by deploying an internal integrator. These examples illustrate an explanatory strategy that is pervasive in computational cognitive science. I call the strategy *function-theoretic explanation* and the mathematical characterization that is central to it *function-theoretic characterization* (hereafter FT).[9] Theories employing the strategy explain a cognitive capacity by appeal to an independently well-understood mathematical function under which the physical system is subsumed. In other words, what gets computed, according to these computational models, is the value of a mathematical function (e.g. addition, vector subtraction, the Laplacean of a Gaussian, a fast Fourier transform) for certain arguments for which the function is defined. For present purposes we can take functions to be mappings from sets (the arguments of the function) to sets (its values). 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. Hence, the FT characterization specifies a kind of content – *mathematical* content – and this content is essential to the computational characterization of the mechanism. If the mechanism computed a different mathematical function, and hence was assigned different mathematical contents, it would be a different computational mechanism.

The mathematical functions deployed in computational models are typically well understood independently of their use in such models. Laplacean of Gaussian filters, vector subtraction, fast Fourier transforms, and so on, are standard items in the applied mathematician's toolbox. An FT description provides an abstract, domain-general, environment-neutral characterization of a mechanism. It prescinds not only from the cognitive capacity that is the explanatory target of the theory (vision, motor control, etc.) but also from the environment in which the capacity is normally exercised.

What I will call the *computational theory proper* comprises a specification of (i) the mathematical function(s) computed by the device (the FT characterization), (ii) the specific algorithms involved in the computation of the function(s), (iii) the representational structures that the algorithms maintain, and (iv) the computational processes defined over these structures. I shall call elements (i)–(iv) the *computational component* of the theory proper. These core elements provide an environment-independent characterization of the device. They 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 to exercise any cognitive capacity at all. Of course, the theorist must explain how computing the value of the mathematical 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 an 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 they occur at different scales, typically correspond to physically significant boundaries – changes in depth, surface orientation, illumination, or reflectance – in the scene. Thus the 'theory proper' will also include (v) 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 in the world. I shall call element (v) the *ecological component* of the computational theory proper. Together these five elements of the theory proper suffice to explain the subject's manifest cognitive capacity.

Cognitive contents

So far nothing has been said about domain-specific representational content. In general, the inputs and outputs of computational mechanisms are characterized not only in abstract terms, as the arguments and values of the specified mathematical function; they are typically also characterized as representing properties or objects relevant to the cognitive capacity to be explained. I call such contents *cognitive contents*. In ascribing cognitive contents the theorist may look for a distal causal antecedent of an internal structure's tokening, or a homomorphism between distal and internal elements, but the search is constrained primarily by the cognitive capacity that the theory is developed to explain. Vision theorists will look to properties that can structure the light in appropriate ways; thus they construe the states and structures they posit as representing light intensity values, changes in light intensity, and further downstream, changes in depth and surface orientation. Theorists of motor control construe the structures they posit as representing positions of objects in nearby space and changes in body joint angles. And the assignment of task-specific cognitive contents will be justified only if the theorist can explain how the posited structures are used by the system in ways that facilitate the cognitive capacity in question.

Cognitive contents, I will argue, are not part of 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 relevant cognitive capacity, though, as noted above, the five elements of the theory proper are strictly speaking sufficient to explain the system's success (and occasional failure). Cognitive contents are best construed as an *intentional gloss* on a computational theory. The primary function of an intentional gloss is to illustrate, in a perspicuous and concise way, how the computational/mathematical theory addresses the intentionally characterized phenomena with which we began and which it is the job of the theory to explain. Cognitive content is the 'connective tissue' linking the sub-personal mathematical capacities posited in the theory and the manifest personal-level capacity that is the theory's explanatory target.

Let me spell out how this works in practice by focusing more closely on the early vision example, although the strategy is general. The computation of the Laplacean of a Gaussian is, of course, presumed to be physically realized in the brain; accordingly, Marr's theory specifies a structure – EDGE[10] – that is the output of this processing. Why would a vision theorist call the structure "EDGE"? Since the structure is individuated by its role in processing, the theorist could have highlighted aspects of its shape, as Marr did for BLOB and BAR, or assigned it an arbitrary name, such as "INTERNAL STRUCTURE #17". Calling the structure "EDGE" highlights its role in the complex process whereby the subject ultimately comes to recover the three-dimensional layout of the scene. So the structure, the output of the processes that

compute the Laplacean of a Gaussian, is glossed in commonsense terms as *EDGE*. To say that the structure *represents* edges is 'shorthand' for the facts that constitute the ecological component of the theory, typically facts about robust covariations between tokenings of the structure and distal property instantiations under normal environmental conditions. These facts explain the organism's visual capacity, and they say nothing about representation.

Recall Chomsky's claim that characterizing an internal structure as 'representing an edge' or 'representing a noun phrase' is simply a convenient way of sorting structures into kinds determined by their role in processing. Chomsky is right that the posited structures are individuated by their computational roles, but wrong to conclude that content serves no legitimate function. The intentional gloss, in assigning contents appropriate to the relevant cognitive domain, shows that the theory addresses its explanatory target, a capacity which is often characterized, pretheoretically by commonsense, in intentional terms (for example, *seeing what is where*).

In addition to the explanatory context – the cognitive capacity to be explained – various pragmatic considerations play a role in determining an appropriate intentional gloss. Given their role in explanation, candidates for cognitive content must be salient or tractable. The structure EDGE represents a change in depth, surface orientation, illumination, or reflectance, but if the distal causes of a structure's tokening are too disjunctive the theorist may decide to assign a proximal content to the structure,[11] motivated in part by a desire to help us (that is, theorists and students of vision) keep track of what the mechanism is doing at a given point in the process.

We can see the extent to which pragmatic considerations figure in the ascription of content by revisiting some of the problems encountered by tracking theories in their attempt to specify a naturalistic content-determining relation. Far from adhering to the strict program imposed by the naturalistic constraint, as understood by tracking theorists, the computational theorist, in assigning content to posited internal structures, *selects* from all the information in the signal what is relevant for the cognitive capacity to be explained and specifies it in a way that is salient for explanatory purposes. Typically, pragmatic considerations will privilege a distal cause (the cat) over a proximal cause (cat-like patterns in the retinal image). Recall the dispute among teleosemanticists about whether the frog's internal state represents *fly* or *frog food* or *small dark moving thing*. The dispute is unlikely to be settled without reference to specific explanatory concerns. If the goal of the theoretical project is to explain the frog's role in its environmental niche, then *fly* content might be privileged. Alternatively, if the goal is to explain how the frog's visual mechanisms work, then *small dark moving thing* might be preferable. In other words, explanatory focus resolves indeterminacy. Turning to Quinean indeterminacy, the ontology implicit in public language privileges *fly* over *fly stage*. But none of these content choices are naturalistically motivated – the naturalistic constraint prohibits appeal to specific explanatory interests or to public meaning.

So we see, then, a *second* function of representational content: to characterize posited internal structures in a way that makes perspicuous their causal role in a process that typically extends into the environment. The content ascription *selects* what is salient in a complex causal process, given specific explanatory concerns. The upshot is quite a different take on the widely accepted view that the content of an internal state or structure *causally explains* the role that the state plays in cognitive processing.[12] This view puts the explanatory cart before the horse. A content ascription captures a salient part of the causal nexus in which the state is embedded. So, for example, construing the frog's internal state as representing *fly* emphasizes the causes of its tokening in the frog's normal ecological niche (its production); construing it as representing *frog food* emphasizes downstream nutritional effects of its tokening (its consumption). Thus it is no surprise that

content *looks* to be causally explanatory – one of its jobs is to characterize internal structures/states in a way that makes perspicuous their causal role in a cognitive process, again, given specific explanatory concerns. But content itself doesn't causally explain anything.

It is time to take stock. The view of content sketched here rejects the two central commitments of tracking theories: (1) Mental representations have their contents *essentially*; and (2) Content is determined by a privileged naturalistic relation holding between the state/structure and the object or property that it is about.

It is typically cognitive (domain-specific) contents that tracking theories take to be both essential to explanations of cognitive capacities and determined by a privileged naturalistic relation. I have argued that the structures posited by computational theories 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 posited structures would be assigned different cognitive contents. If the subject's environment were different, so that the use of these structures by the device 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. Moreover, since pragmatic considerations typically *do* play a role in determining cognitive contents, these contents are not determined by a naturalistic relation.

Turning to mathematical contents: whatever the ontological status of mathematical objects, it is unlikely that any naturalistic relation holds between the structures posited in the theory and (just) the mathematical objects specified by the FT characterization. Nonetheless, mathematical content *is* essential. The various scenarios discussed above would not affect the attribution of mathematical content, because the FT characterization is a canonical specification of what the device does. To characterize the device as computing a mathematical function *just is* to interpret its inputs and outputs as representing the arguments and values of the function respectively; if the FT characterization is essential, as I have argued, then so is the mathematical content that it determines.

Revisiting the adequacy conditions

I shall conclude by considering this deflationary account of content in light of the adequacy conditions for a theory of content for computational neuroscience.

Condition (1) requires that the account provide the basis for the attribution of *determinate* contents to computational states or structures. The deflationary theory does better in this respect than tracking theories, all of which have trouble grounding determinate content in a naturalistic relation. Once the role of specific explanatory interests and other pragmatic factors in content attribution is fully appreciated, determinacy is to be expected.

Condition (2) requires that the account allow for the possibility of *mis*representation. There is no mystery about how misrepresentation arises in the deflationary account. Cognitive contents are ascribed to internal structures on the basis of the cognitive capacity to be explained, what is happening in the subject's normal environment when the structures are tokened, and various pragmatic considerations discussed above. Normally, the structure is tokened when and only when the specified external condition obtains. But occasionally something goes wrong. In low light, a shadow may be mistaken for an edge. In an Ames room at Disney World, where the light is systematically distorted, the subject will misjudge the character of the local space. In such circumstances, the structure whose cognitive content is *edge* is tokened in response to a shadow or some other distal feature, and the mechanism

misrepresents a shadow as an edge. The 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. Lest one think that a tracking theorist could avail herself of a similar story about misrepresentation, keep in mind that the structures have their (determinate) cognitive contents only in the gloss, where various pragmatic considerations provide the additional constraints necessary to support an attribution of misrepresentation. Misrepresentation, like veridical representation, is confined to the intentional gloss.[13]

Condition (3) requires that the account be naturalistic. At first blush, it may seem that the appeal to explanatory and other pragmatic considerations in the determination of cognitive content compromises the deflationary account's naturalistic credentials. That isn't so, because the pragmatic elements and the contents they determine are 'quarantined' in the intentional gloss, to use Mark Sprevak's (2013) apt expression. The theory proper of a cognitive capacity ((i)–(v) above) does not traffic in ordinary (i.e. cognitive domain-specific) representational contents. The theory proper provides a full description of the capacity sufficient to explain the organism's success at the cognitive task; the intentional gloss serves the various heuristic purposes described above.

Recall that the primary motivation for the naturalistic constraint is the conviction that intentionality is not fundamental, and the hope among materialistically minded theorists of cognition that computationalism will contribute to a naturalistic reduction. Specifying non-intentional and non-semantic sufficient conditions for an internal state's having its determinate content – the project that tracking theorists have set for themselves – is only one way that a reduction might be accomplished, and not a particularly promising way if, as I have argued, content attribution in computational practice is rife with pragmatic elements. But insofar as the deflationary account sketched here is an accurate representation of that practice, computational neuroscience is making some progress toward a naturalistic reduction of intentionality. I don't want to overstate the point: computational theories appeal to unreduced mathematical content. But a well-confirmed computational theory of a cognitive capacity that included an account of how the mechanism is realized in neural structures would be a significant step toward a reductive explanation of intentionality in that cognitive domain. States and structures that are characterized in the theory in terms of their computational role have meaning and truth conditions only in the intentional gloss, where they are used to show that the theory addresses the phenomenon for which we sought an explanation.

One of Chomsky's motivations for eliminating representational content is the desire to purge the cognitive sciences of normative and intentional notions – such talk as 'solving a problem', 'making a mistake', 'misrepresenting' – which he thinks reflect our parochial interests, and hence have no place in legitimate science. But such austerity is neither necessary nor appropriate. The project, after all, is to understand our own mentality. The intentional gloss characterizes computational processes in ways congruent with our commonsense understanding of ourselves, ways that the theory itself eschews. It fills a kind of explanatory gap between the scientific and the manifest image, to put the point in Wilfrid Sellars' (1963) terms.

Finally, condition (4) requires that the account conform to actual practice in computational cognitive science. The deflationary account improves on its competitors in two significant respects: (1) it recognizes the role played in computational models by a mathematical characterization of a mechanism, and hence the attribution of mathematical content, and (2) it explicitly acknowledges the role of pragmatic considerations in the ascription of ordinary representational content.

Notes

1 The recent resurgence of panpsychism notwithstanding.
2 See Dretske (1981) and Fodor (1990) for the most developed information-theoretic accounts. Further conditions include the requirement that during a privileged learning period only cats cause S-tokenings (Dretske, 1981) or that non-cat caused S-tokenings depend asymmetrically on cat-caused S-tokenings (Fodor, 1990).
3 See Matthen (1988), Papineau (1993), Dretske (1995), Ryder (2004), Neander (2006; 2017), and Shea (2007) for other versions of teleosemantics.
4 See the discussion of the magnetosome in Dretske (1986) and Millikan (1989).
5 See Quine (1960).
6 The expression is from Kriegal (2013).
7 See Chomsky (1995; 2000). For another eliminativist view see Stich (1983).
8 See Egan (2013) for elaboration of the account sketched here.
9 See Egan (2017) for elaboration of FT explanation.
10 I will use upper case to denote structures whose individuation conditions are given by their roles in processing, i.e. non-semantically.
11 For example, zero-crossings in Marr's theory represent discontinuities in the image.
12 For a sample of the literature promoting this idea see Dretske (1988), Segal and Sober (1991), and Rescorla (2014).
13 The structures characterized abstractly in the theory by the FT specification can also misrepresent. If the mechanism overheats or is exposed to a harmful substance it may fail to compute its normal mathematical function, for example, miscomputing, and hence misrepresenting, the sum of a vector addition as some other value.

References

Chomsky, N. (1995) 'Language and Nature', *Mind*, 104, pp. 1–61.
Chomsky, N. (2000) 'Internalist Explorations', in *New Horizons in the Study of Language and Mind*. Cambridge, UK: Cambridge University Press, pp. 164–194.
Chomsky, N. (2003) 'Reply to Egan', in Antony, L. and Hornstein, N. (eds.) *Chomsky and His Critics*. Oxford: Blackwell, pp. 268–274.
Cummins, R. (1989) *Meaning and Mental Representation*. Cambridge, MA: MIT Press.
Dretske, F. (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
Dretske, F. (1986) 'Misrepresentation', in Bogdan, R. (ed.) *Belief: Form, Content, and Function*. Oxford: Oxford University Press, pp. 17–36.
Dretske, F. (1988) *Explaining Behavior*. Cambridge, MA: MIT Press.
Dretske, F. (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.
Egan, F. (2013) 'How to Think about Mental Content', *Philosophical Studies*, 170, pp. 115–135.
Egan, F. (2017) 'Function-Theoretic Explanation and the Search for Neural Mechanisms', in Kaplan, D.M. (ed.) *Explanation and Integration in Mind and Brain Science*. Oxford: Oxford University Press, pp. 145–163.
Field, H. (1975) 'Conventionalism and Instrumentalism in Semantics', *Nous*, 9, pp. 375–405.
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, pp. 89–136.
Horgan, T. and Graham, G. (2012) 'Phenomenal Intentionality and Content Determinacy', in Schantz, R. (ed.) *Prospects for Meaning*. Boston, MA: De Gruyter, pp. 321–344.
Kriegel, U. (2013) 'The Phenomonal Intentionality Research Program', in Kriegel, U. (ed.) *Phenomenal Intentionality*. New York, NY: Oxford University Press, pp. 1–26.
Marr, D. (1982) *Vision*. New York, NY: Freeman.
Matthen, M. (1988) 'Biological Functions and Perceptual Content', *Journal of Philosophy*, 85, pp. 5–27.
Millikan, R. (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Millikan, R. (1989) 'Biosemantics', *Journal of Philosophy*, 86, pp. 281–297.
Neander, K. (2006) 'Content for Cognitive Science', in Papineau, D. and McDonald, G. (eds.) *Teleosemantics*. Oxford: Oxford University Press, pp. 140–159.
Neander, K. (2017) *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

Papineau, D. (1993) *Philosophical Naturalism*. Oxford: Blackwell.
Quine, W.V. (1960) *Word and Object*. Cambridge, MA: MIT Press.
Ramsey, W. (2007) *Representation Reconsidered*. Cambridge, UK: Cambridge University Press.
Rescorla, M. (2014) 'The Causal Relevance of Content to Computation', *Philosophy and Phenomenological Research*, 88, pp. 140–159.
Ryder, D. (2004) 'SINBAD Neurosemantics: A Theory of Mental Representation', *Mind and Language*, 19, pp. 211–240.
Segal, G. and Sober, E. (1991) 'The Causal Relevance of Content', *Philosophical Studies*, 63, pp. 1–30.
Sellars, W. (1963) 'Philosophy and the Scientific Image of Man', in *Science, Perception, and Reality*. New York, NY: Humanities Press.
Seung, S.H. (1996) 'How the Brain Keeps the Eyes Still', *Proceedings of the National Academy of Science USA*, 93, pp. 13339–13344.
Seung, S.H. (1998) 'Continuous Attractors and Oculomotor Control', *Neural Networks*, 11, pp. 1253–1258.
Seung, S.H. et al. (2000) 'Stability of the Memory of Eye Position in a Recurrent Network of Conductance-based Model Neurons', *Neuron*, 26, pp. 259–271.
Shadmehr, R. and Wise, S. (2005) *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA: MIT Press.
Shagrir, O. (2012) 'Structural Representations and the Brain', *British Journal for the Philosophy of Science*, 63, pp. 519–545.
Shea, N. (2007) 'Consumers Need Information: Supplementing Teleosemantics with an Input Condition', *Philosophy and Phenomenological Research*, 75, pp. 404–435.
Sprevak, M. (2013) 'Fictionalism about Neural Representations', *The Monist*, 96, pp. 539–560.
Stich, S. (1983) *From Folk Psychology to Cognitive Science: The Case against Belief*. Cambridge, MA: MIT Press.