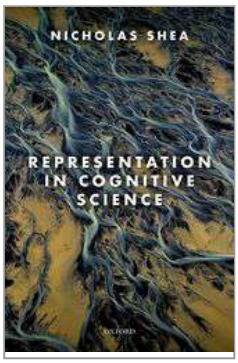

Representation in Cognitive Science
Nicholas Shea

Print publication date: 2018
Print ISBN-13: 9780198812883
Published to Oxford Scholarship Online: October 2018
DOI: 10.1093/oso/9780198812883.001.0001

How Content Explains

Nicholas Shea

DOI:10.1093/oso/9780198812883.003.0008

Abstract and Keywords

The varitel accounts of content allow us to see how the practice of representational explanation works and why content has an explanatory role to play. They establish the causal-explanatory relevance of semantic properties and are neutral about causal efficacy. Exploitable relations give the accounts an advantage over views based only on outputs. Content does valuable explanatory work in areas beyond psychology, but it need not be explanatorily valuable in every case. The varitel accounts illuminate why there should be a tight connection between content and the circumstances in which a representation develops. The accounts have some epistemological consequences. Representations at the personal level are different in a variety of ways that are relevant to content determination. Naturalizing personal-level content thus becomes a tractable research programme. Most importantly, varitel semantics offers a naturalistic account of the content of representations in the brain and other subpersonal representational systems.

Keywords: representational explanation, explanatory purchase, epiphenomenalism, causal efficacy, exploitable relation, representational development, personal level, metacognition, consciousness, naturalism

8.1 Introduction

This chapter offers some theoretical reflections on the accounts of content presented in previous chapters. First, in this section I briefly reiterate some of the distinctive features of my view.

One unusual feature of the book is that it devotes so much space to a series of detailed case studies. The aim of that was to understand how representation is used right across the cognitive sciences. Where it leads is pluralism. With two exploitable relations and a variety of task functions, my ‘theory’ of content is in fact a collection of different theories. Pluralism has been suggested before, but not until now worked up into a collection of detailed, mutually compatible accounts. My approach is unusual in focusing exclusively on subpersonal cases, and in the extent to which I’m interested in neural representation. Renouncing representation consumers is a new way to develop teleosemantics. We can get the benefits of representationalism—the explanatory benefits that flow from having vehicles of content—without consumers, and also without collapsing into an instrumentalist or ascriptionist view.

(p.198) Being careful about the value of RTM led to the view that content arises from convergence between task functions, internal processes and exploitable relations: it arises when internal processing over vehicles standing in exploitable relations to the environment implements an algorithm for performing the organism’s task functions. The idea of vehicles being processed in virtue of non-semantic properties in ways that respect their contents is of course not new; nor is the idea of exploitable relations (Godfrey-Smith 2006). However, the way varitel semantics puts these ideas together as the basis for content determination is distinctive. The focus on explaining the explanatory purchase of representational content is also a new emphasis, leading to an original proposal about why the world affords us the representational scheme of explanation—because of there being a natural cluster in which stabilizing processes go together with robust outcomes and internal mechanisms for producing them.

Section 8.2 returns to the question of content’s explanatory purchase and shows that the accounts in Chapters 3–7 do deliver on the promissory note in Chapters 1 and 2. The varitel accounts allow us to see how the practice of representational

explanation works and why content has an explanatory role to play. Section 8.3 looks at the causal efficacy of semantic properties, on the varitel view. Section 8.4 asks whether exploitable relations are doing any substantive work in the accounts, or whether they are dispensable in favour of an output-only approach to content. Section 8.5 asks how far varitel semantics extends, whether it applies to cases where content is un-explanatory, and whether it applies too widely. Section 8.6 remarks on the tight connection that exists between content determination and the circumstances in which a representational capacity develops, suggesting that this fits well with other issues in the literature. Section 8.7 goes through a list of clarifications and qualifications that couldn't easily be dealt with earlier. Section 8.8 draws out some epistemological consequences from our (metaphysical) accounts of content determination. Finally, section 8.9 suggests some ways that differences at the personal level may turn out to be relevant to content determination there.

8.2 How Content Explains

(a) Explanatory traction in varitel semantics

The varitel framework was motivated by the desideratum that we should be able to explain how content-based explanation works. The answer sketched in Chapter 2 started with the idea that contents have real vehicles because content explanation is partly concerned with explaining how a system manages to generate appropriate behaviour. Contents are externalist because the patterns of behaviour to be explained are world-involving: achieving distal effects in the world by reacting to distal objects and properties. The extrinsic properties that are relevant to explaining how the organism does that are exploitable relations that vehicles of content stand in to features of its environment. These externalist properties are suited to explaining how internal **(p.199)** processing implements an algorithm for carrying out an organism's distal functions. That was all effectively a promissory note: if I can devise a theory of content that fits within the framework, then that should allow us to see how representational contents are suited to explaining behaviour.

Now that I have pinned my colours to the mast and set out a series of accounts of content, the time has come to assess whether the accounts deliver. Do they allow us to see how contents explain behaviour? In particular, do they throw light on the characteristic explanatory grammar of representational explanation (§2.2): that correct representation explains successful behaviour, and misrepresentation failure?

That explanatory grammar arises naturally from my accounts of content. Take the analogue magnitude system as an example (§4.6a). Consider a primate trained to choose between two sets of objects, being rewarded for selecting the more numerous collection. The training has used and tuned the animal's analogue magnitude subsystem, giving the animal a disposition to pick the thing in the world corresponding to whichever analogue magnitude register in its

parietal cortex is more active. The training also gives rise to a standard of success and failure for the animal's actions in these contexts. When presented with two buckets containing collections of objects, picking the more numerous collection is a successful behaviour. Behaving in this way is a task function of the organism.

This task function not only constitutes a standard for success and failure of behaviour. It also specifies a mapping from distal situations (e.g. objects in buckets) to distal outcomes (e.g. picking up the fuller bucket), a mapping mediated by the animal. The monkey in its environment is disposed to instantiate this input–output mapping (at least approximately). There may be more than one algorithm that would generate this input–output performance, but there is a fact of the matter about which algorithm is at work inside the monkey. The algorithm is specified in world-involving terms: individuate the objects in one bucket, add up the total number of objects on that side, compare the total number of objects on each side, choose the collection with the largest number of objects. Internal processes inside the monkey are specified in intrinsic terms: patterns of neural firing here cause patterns of neural firing there cause … cause bodily movements. What makes it the case that this internal process implements an algorithm for performing the task is the correlational information carried by each component. Each component correlates with a distal fact called for by the algorithm (e.g. the number of objects on one side). Furthermore, the way internal processes operate over these vehicles implements the transitions called for by the algorithm. Thus, having content constituted by the convergence between exploitable relations and task functions, in the ways set out in previous chapters, both implies that there is a difference between successful and unsuccessful behaviour, and also makes for contents suited to explaining how an organism responds to distal facts in its environment so as to produce the distal outcomes which count as successes. Correlatively, misrepresentation by an internal component will explain unsuccessful behaviour.

(p.200) I argued in Chapter 3 that this whole explanatory practice gets traction because of a deep fact about the world we live in. Things produced by natural selection tend to be disposed to produce outcomes robustly, because when an outcome is the target of selection, evolution can find ways for it to be achieved more robustly. Evolution's greatest robustness trick is the organism itself: a complex system that differentiates itself from its environment and continually maintains itself in a state that is out of equilibrium with the environment. Organisms produce the conditions needed for their own persistence, including by modifying their dispositions to achieve that end through learning. Furthermore, learning itself is a stabilizing process by which an individual can come to produce outcomes more robustly. So it is no accident that the biological realm is full of goal-directedness in the Aristotelian sense: robustly produced

outcomes that have been stabilized by natural selection, learning, or contributing to the persistence of an organism.

One very common way—although by no means the only way—that organisms produce outcomes robustly is by having an internal mechanism that keeps track of aspects of the environment and thus implements an algorithm for performing the input–output mapping that has been the target of stabilization. That is, they do it by using representations. Nature has given us a widely implemented cluster: stabilization and robustness achieved by internal workings bearing exploitable relations. Representational explanations take advantage of the generalizations and inductions afforded by this natural cluster. Artefacts that humans have designed, like control mechanisms and computers, also often fall into the cluster, and for the same reasons. Tying representation to this cluster, rather than something more liberal, is the source of its inductive power: UE information and UE structural correspondence get explanatory purchase from the fact that their instantiation goes along with a cluster of other properties.

(b) Non-semantic causal description?

In addition, varitel semantics allows us to answer a familiar challenge to the status of representational explanation (§2.3). Isn't there an entirely non-semantic causal description of how any organism or system will react to inputs, undergo internal changes, and produce outputs? Realists about representational vehicles are committed to there being a non-semantic (vehicle-based) causal description at the same level as the semantic description.[1] If we can give a non-semantic causal description of the internal operation and outputs of a system moment-by-moment, what does representational content add?[2]

(p.201) Advocates of the explanatory force of representational explanation can point to all the successes of representation-based psychology—a huge body of work containing rich generalizations about representations in general, and especially about specific kinds of representation (motor programs, reward prediction errors, analogue magnitude representations, etc., etc.). But the vehicle-based challenge threatens to undermine the seemingly obvious explanatory traction of psychology by showing that it has no autonomy from a non-semantic element in its foundations. The aim of this subsection is not to catalogue the rich generalizations that give representational explanation its explanatory potency—we can turn to any psychology textbook for that. It is to show how varitel semantics answers the challenge. Varitel semantics has a feature which allows representationalism's commitment to non-semantic vehicles to be compatible with content having distinctive explanatory purchase.

To return to Ramsey's example (§2.2), consider the firing mechanism of a rifle (see Figure 2.2). Some theories of content imply that the displacement of the firing pin represents that the user's finger has been pulled back and instructs the cartridge to fire a bullet. If semantic contents were like that, then

representational explanation would march exactly in step with a ‘factorized’ causal chain:

(i) The user’s finger moves backwards.
(ii) The trigger moves backwards.
(iii) The firing pin shoots forwards.
(iv) The charge in the cartridge ignites.
(v) The bullet flies off at speed.

Steps (ii)–(iv) form a causal chain given in terms of intrinsic properties of the rifle. Step (i) is external to the rifle and causes (ii), which is a process intrinsic to the rifle. Step (iv) is also intrinsic to the rifle and causes step (v), which is an outcome external to the rifle. The causal chain is factorized into events in the external environment, on the one hand, and events intrinsic to the rifle, on the other. If movement of the firing pin were to carry semantic content, then there would be another explanation of the rifle’s behaviour: movement of the user’s finger leads to a representation being tokened with the content *the user’s finger has been pulled back, fire a bullet*, which leads to a bullet being fired. That explanation marches exactly in step with the non-semantic explanation above. It is just a semantic relabelling of the process (i) → (iii) → (v).

Varitel accounts of content imply that semantic explanations of behaviour do not march exactly in step with a factorized causal explanation of how an organism **(p.202)** responds to proximal inputs with bodily movements. Task functions are robust outcomes, so the same outcome is produced in response to a range of different proximal inputs. That means that vehicles of content will enter into generalizations that ‘bridge’ across multiple proximal conditions and involve distal states of affairs (see Figure 8.1).[3] Connections like the one between steps (i) and (iii) above, rather than being mediated by a single proximal input (ii), will be mediated by a range of different proximal inputs (ii*), (ii**), etc. There is a distal exploitable relation without a matching proximal exploitable relation. The representational explanation does not then simply march in step with the factorized explanation. Causal steps which show up as different in the factorized explanation are unified in the representational explanation. The representational explanation is picking up on patterns in vehicle–world relations that the factorized explanation would miss. The same is often true at output since outcomes that count as task functions tend to be produced by the organism in a variety of different ways in different circumstances; that is, via a variety of different bodily movements (§3.3 and §3.6). Those are further patterns, captured by the representational explanation, that a factorized explanation would miss.

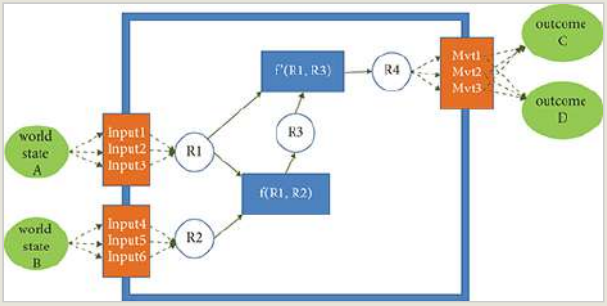

Figure 8.1 A schematic depiction of bridging at input and output.

This bridging means that there are real patterns in organism–world relations—in relations of internal states to distal causes and outcomes—that are treated disjunctively in the factorized explanation. The effect of past processes of stabilization has been to key the organism into the world so that generalizations do not just concern how proximal causes affect the organism and how the organism affects its immediate proximal environment. A purely factorized explanation would miss the distal patterns. Dennett argued that belief-desire explanation picks up on real patterns in the way **(p.203)** agents interact with the world (Dennett 1991). Not only would prediction of agents' behaviour be impossible in practice if we treated them as collections of molecules interacting with other molecules in the environment, but also such a low-level physical description would fail to capture real patterns that exist—that are an interpreter-independent feature of a world full of agents.[4] My approach to content similarly shows how content-based explanation takes advantage of real patterns in the world, patterns that exist because of regularities in biological and physical processes, and which exist irrespective of the existence of observers to notice them.[5]

In the absence of a robust outcome function this argument does not get off the ground. Then a factorized explanation may march exactly in step with the putative representational explanation. We saw that in the case of the rifle. That is not a case where representational content would give any additional explanatory purchase, and indeed varitel semantics does not imply that movement of the rifle firing pin would carry representational content. The same point applies to the magnetotactic bacteria. As the case is standardly described (Dretske 1986, Millikan 1989, Cummins et al. 2006), moving in the direction of oxygen-free water is a stabilized function of the bacterium's behaviour, but not a robust outcome function.[6] Our accounts do not imply that the bacterium or its magnetosome carries representational content. Nor is it a case where content would afford a better explanation than a non-semantic causal description (one involving functions but not representational contents).

We can see bridging at work in our case studies. Consider again a monkey that deploys its analogue magnitude system to choose the more numerous set of objects in a range of different circumstances. At input, analogue magnitude registers in the parietal lobe correlate with the numerosity of distal collections of objects, a correlation that is mediated by a variety of proximal inputs: many different kinds of visual patterns, auditory patterns, etc. At output, the action of

reaching out to touch or grab the more numerous collection is mediated by a variety of different motor outputs in different circumstances. So, there is bridging at input and at output. The rat's system for spatial navigation, using the hippocampus and other areas, also shows bridging. The input sensitivity of place cells bridges across multiple patterns of sensory input (e.g. visual input from different orientations). Behavioural output also exemplifies **(p.204)** bridging. The rat can reach a baited or rewarded location from a range of different starting positions by a range of different routes, relying on structural correspondence in the way described in §5.3.

If we were just to look at the rat's limb movements taken in isolation, without considering how those bodily movements produced locomotion or where the animal was in space, we would see a variety of nearly uninterpretable movements of different limbs with different speeds and in different directions at different times. It would be like watching the bodily movements of a teenager playing a video game on a smartphone, but without being able to see the screen. The two thumbs move rapidly in seemingly arbitrary ways, with no apparent pattern. Only once the relation of those movements to what is happening on the screen comes into view do they become interpretable. The real patterns are not found in when the thumb moves, but in what happens to the character on the screen, and how that relates to movements made by the game-player, and to her intentions. There are real patterns in what is happening in the player's environment. The thumb movements act as 'bridging' causal intermediates. Similarly, there are very clear patterns in the rat's behaviour if we consider it in relation to its environment. Those patterns are mediated by and generalize—bridge—across a diversity of bodily movements output by the rat.

The content-based generalizations in psychological theories link representation with representation, often of specific types, representation with neural substrate, and representation with world. These are by no means all cases of bridging. What bridging does is to show how content-based explanation can break free from non-semantic vehicle-based explanation, allowing the rich and detailed theories of psychology and cognitive neuroscience to have their own explanatory purchase.

Bridging exemplifies one explanatory virtue, generality. It groups together things that would otherwise be classified as different. But the previous subsection (§8.2a) argued for a seemingly contradictory virtue, specificity. An account that made content very liberal would be problematic and an advantage of varitel semantics was that its contents only arise when a special cluster of properties occurs (which it often does, for natural reasons). There is in fact no contradiction because the advantage claimed for the cluster is its inductive potential. Finding UE information or UE structural correspondence implies many other things about the system in question. In fact, bridging relies on inductive power as well. Content is based on a bridged relation but shows up in

generalizations with other properties. With explanation, there is always a balance between properties that apply widely enough to support good generalizations and properties that are specific enough that they support rich inductions. Our accounts show how content properties strike a balance that gives them genuine explanatory traction.

(c) Doing without talk of representation

Another kind of challenge is faced by any theory of content which offers non-semantic, non-mental, non-normative conditions under which representational content arises. Suppose it is granted that correlation, correspondence, and function come together in **(p.205)** the special way I have claimed, so that their convergence affords generalizations and inductions. However, the objection runs, we can recognize all that without ever mentioning representation or content.[7] Why can't we do all the explanatory work directly in terms of correlation, correspondence, and function? Indeed, the accounts of content set out above give us precisely the tools we need in order to do just that.

There are two versions of this challenge, which should be answered in different ways. The first rival to content-based explanation only helps itself to the resources that appear in my definitions—correlation, correspondence, robustness, stabilization, etc.—and does not advert to the fact that these properties tend to come together in the packages I have pointed to. Instead of contents, they offer explanations directly in terms of the properties in the explanatory base. They replace content-based generalization with much more fine-grained explanations. The trouble with these views is their complexity. More complex properties are generally less good candidates for explanation. Furthermore, it is not clear why relational properties like correlating or producing outcomes robustly get any explanatory purchase, if the motivation I have offered in terms of a natural clustering of properties is absent. Such a rival would also miss the inductive potential that exists—the fact that robustness and stabilization do tend to converge (so as to constitute task functions) and the fact that internal workings, correlation, correspondence, and task function do tend to converge in regular ways (which is what I say constitutes content).

The second version of the challenge recognizes the clusters set out in our accounts of content, but objects to these being identified as representations. We can do all the same explanatory work by recognizing that there are these real clusters and making generalizations and performing inductions over instances of them. This supposed challenge is not really a challenge at all, because it concedes everything we need, leaving only a dispute about the appropriateness of the label 'representation'. A distinctive style of representation-based explanation that our theories of content need to explain is that correct representation explains successful behaviour and misrepresentation explains failure; that is a form of explanation where the obtaining or otherwise of facts that are distal to the organism or system make a difference to explaining its

behaviour. Explanations like that are part of what made representational content puzzling, even mysterious. Recognizing that the clusters I point to are real and important features of the natural world, and accepting that they underpin world-involving explanations of this kind, just is to accept that properties of the kind I have characterized do exist, and do explain behaviour in the way I have claimed.

(d) Other views about the explanatory purchase of content

Now that I have laid out my view about the explanatory purchase of representational content, I will compare it briefly to some other views in the literature. William Ramsey **(p.206)** identifies two broad ways in which representational properties have been argued to earn their explanatory keep (Ramsey 1997, p. 37). First, they may have heuristic value in allowing us to think about a system in a useful way. They can play that role despite having no causal relevance to how the system operates.[8] Some adopt this view about the computational theory of mind: syntax does all the causal work, but semantics allows us to see why a system's syntactic processes are suitable for performing certain computations. The second option is that contents are causally explanatory of behaviour; for example, because they are a structuring cause of dispositions to behaviour, as argued by Dretske (1988). (A further view is that representational content does not earn its explanatory keep at all, and so should be abandoned, e.g. Stich 1983.)

Frances Egan is in the first camp (Egan 2014). Content for Egan (her 'cognitive content') is useful because it allows the theorist to understand how a computational system can perform a cognitive task; for example, the task of seeing what is in the nearby environment. Different contents will be useful for explaining how the same computational system performs different cognitive tasks. In each case, content is just a gloss, useful to the theorist. Oron Shagrir falls in the same camp, but with a more realist take on content (Shagrir 2006). For him too, theorists adopt the representational approach in order to explain semantic tasks performed by a system. These views are similar in one way to Tyler Burge's theory, since Burge takes the target of representational explanation to be the capacity for perception and the computations and transformations involved in perception (Burge 2010). In all these cases the task that calls for explanation is already stated in semantic terms. Contentful states enable us to understand how an organism can perform a cognitive, semantic task.

Dretske (1988) exemplifies the second camp, since he argues that contents are causally explanatory. Content arises when an internal state R has been recruited as a cause of behavioural output M in virtue of the fact that R carries information about[9] condition C. The fact that information has converged with learning in the past—the fact encapsulated by the existence of content—causally

explains why the organism is configured as it is today, and hence forms part of an explanation of why it produces behaviour M on a particular occasion.[10]

Another way that contents can have causal relevance arises specifically for conceptual contents. Possessing concepts can explain why certain capacities of an organism are systematically related. The compositionality of concepts is then the source of a special **(p.207)** explanatory value of representational locutions, because it allows us to explain the systematicity of cognitive capacities, although here too non-semantic vehicle properties are rivals for causal relevance (Camp 2009, Fodor 1987b).

Varitel semantics partakes of elements of both camps. Within the second camp, like Dretske I rely on considerations about why the system has been configured the way it is and behaves the way it does.[11] In varitel semantics content arises when an organism has a disposition to produce certain outcomes because exploitable relations converge with stabilizing processes that have operated on those outcomes in the past. Ramsey (2007, pp. 132–40) objects to Dretske's theory on the basis that it over-generates: not all cases where indicator properties are a structuring cause of behavioural outputs should count as representational. My accounts have stronger requirements than Dretske's and so avoid this liberality objection. Nevertheless, part of the explanatory power of content, on my view, traces to the kind of causal process identified by Dretske.

My accounts share with the first camp the view that contents are useful because they allow us to see why an organism's internal workings are suited to performing certain tasks. Unlike Egan and Shagrir, I characterize those tasks in non-semantic terms in the first instance. Mine are not cognitive tasks but mappings from worldly conditions to distal outputs (outputs which qualify as task functions). But like Shagrir, contents for me are partly a matter of how an organism can perform the computations needed to produce appropriate outputs in appropriate circumstances. I reject Egan's contention that contents are merely a theorist's gloss, with a different gloss being appropriate when a system is located in different contexts. The context in which a system is operating is a property of the system, just as much as its intrinsic properties are, and I take that context to have a content-determining rather than merely a pragmatic role.

Since concepts have not been part of our investigation, I have said little about the role of representations in explaining systematicity. Nevertheless, we saw in §6.3 that the representations in some of our case studies do have semantically significant constituent structure. Even without semantically significant structure at the level of representational vehicles, the division of an organism's internal workings into a series of algorithmic steps has some of the flavour of systematicity (§5.7a, §6.3).[12] In both cases, facts about vehicles and how they interact explain certain systematic patterns in the organism's behaviour. The point I have laboured about vehicle realism and internal workings (§1.3, §2.5,

§3.2, §8.2a) is in fact a generalization of others' observations about systematicity of structured representations and the explanatory purchase of content (Fodor 1975, 1987b; Fodor and Pylyshyn 1988; Camp 2009).

In short, varitel semantics can adopt sound arguments about explanatory purchase drawn from both of the broad camps identified by Ramsey (1997, p. 37).

(p.208) 8.3 Causal Efficacy of Semantic Properties

The previous section showed why representational explanations are partly autonomous from purely vehicle-based explanations, and hence why semantic properties can afford us some additional purchase in explaining behaviour. But are content properties causally efficacious, or simply explanatorily relevant? Jackson and Pettit distinguish between process explanation and program explanation (Jackson and Pettit 1988, 1990). They argue that the properties cited in a program explanation can be explanatorily relevant without being causally efficacious—when it is the properties cited in the process explanation that do the real causing. For example, the squareness of a wooden peg explains why it won't fit into a round hole of the same surface area. However, the causally efficacious properties, it is argued, do not involve squareness or roundness, but other physical properties of the material of the peg and hole.

Jackson and Pettit introduce their distinction in order to save the explanatory relevance of broad contents, and it is equally applicable to our case. The vehicle-based explanation of behaviour may be telling us where the real causal work is going on, but nevertheless content properties can be explanatory. This casts varitel accounts of content as picking up on opportunities for program explanation. When the explanandum occurs in a range of cases, and so is more general than any specific causal process, a program explanation tells us that what matters for achieving a result is that some relational state of affairs obtains, irrespective of which particular state of affairs causes it to obtain. Program explanations 'tell us about the range of states that do or would produce the result without telling us which state in fact did the job' (Jackson and Pettit 1988, p. 396).

If contents are fixed in the way I have claimed, then the argument in the previous section shows why semantic properties figure in program explanations, hence are explanatory of behaviour. That they are not causally efficacious is a further claim. One challenge is faced by all special science properties, the challenge that the 'real' causal work is being done at some more fundamental level. But if we thought that the real causal work was captured by the factorized, vehicle-involving explanations mentioned above, then we would face a further challenge: vehicles are realized by physical properties (e.g. neural firings). Surely the vehicle-based description is only a program explanation, with the causal work really going on at a more fundamental neurophysiological level?

But, of course, the same move arises there, with a molecular, chemical, and electrical explanation threatening to displace the causal efficacy of neural depolarizations. The descent continues, until we reach the level, if there is one, of the most fundamental physics (where, incidentally, it is arguable that causation does not figure at all).

An alternative is that causation—'real' causal efficacy—is found at more than one level in this hierarchy. There may be no good reason to think that causal processes that are connected by relations of constitution should exclude each other (Bennett 2003). Thus, even if it is granted that some patterns of explanation deal in explanatory relevance rather than causal efficacy, it would be hasty to conclude that content-based explanations are not a locus of causal efficacy.

(p.209) If causal efficacy is a tenable position for some special science properties, there are further obstacles to making that case for content properties. Contents are partly historically based, and are partly tied to the kinds of effects they are called on to explain (Shea 2007b). I have argued that these are not obstacles to explanatory purchase, but it is a further step to show that these features are compatible with the causal efficacy of content. We would have to see that the world-involving generalizations that contents are involved in support counterfactuals and interventions in the right way, and possibly that they count as figuring in a causal process (on process-involving views of causation). We would also need to establish that the relation between content properties and the vehicle-based story does not generate causal exclusion, especially in the light of the fact that some of the representation-to-representation transitions (or inferences) featuring in the semantic-level explanation have an exact parallel in the syntactic explanation. Nevertheless, it could be that the forms of generalization, counterfactual-dependence, and underlying process that embed these transitions in world-involving patterns amount to set of genuinely causal relations.

These are large issues, well beyond the scope of this book. For now I want to remain neutral about whether contents are causally efficacious. Varitel semantics allows us to see why semantic properties figure in program explanations, so I rest with the positive claim that content properties are at least explanatorily relevant.

8.4 Why Require Exploitable Relations?

My accounts require that representations bear exploitable relations to the things they represent. More carefully, exploitable relations have to be in place when behaviour is stabilized (when the stabilizing process, which makes it the case that the system has a function that is partly constitutive of content, is operative). The leading teleosemantic theories of Millikan and Papineau eschew any requirement of this kind. Contents are an output-only matter, determined by the

functions of behaviour prompted by a representation, and by evolutionarily normal conditions for performing those functions. Are exploitable relations dispensable? It is not in dispute that exploitable relations will normally be in place during stabilization. For example, dances of incoming honeybees would have correlated with nectar location at the time of selection. But do correlations need to figure in the content-constituting story?

My account agrees with teleosemantics that functions are an essential part of the story. They furnish an explanandum to which representational explanation is directed, namely successfully producing distal effects in the environment, or failing to do so. Teleosemantics says that contents are a matter of functions and conditions under which functions will be successfully performed. Varitel semantics goes further. It says that content is partly a matter of explaining how a system achieves its functions—of how its internal workings implement an algorithm for performing its functions. That is why exploitable relations get into the picture.

(p.210) Without exploitable relations, representational contents are effectively a way of typing internal states by patterns in the outputs they (historically) produce. Adding exploitable relations on the input side furnishes a reason why, when the representation is tokened, it is likely that its correctness condition obtains.[13] That will be true at the time of stabilization and, to the extent that things haven’t changed, will also be true of current behaviour. That is, our accounts do not just produce correctness conditions that are a typing of behaviour. They give us a reason to think that those correctness conditions obtain—a stronger reason than teleosemantics is committed to.

That difference gives varitel accounts more predictive power: we can use content to predict what the system will do because, if the environment is stable in relevant respects, representations will continue to correlate or structurally correspond with the world, giving us stronger predictions about the distal results that are likely to flow from outputs produced by the system. That difference is not profound, however, because teleosemantics can appeal to something very similar: an empirical generalization that, where there are contents of the teleosemantic sort, exploitable relations are usually in place. So the predictive advantage of varitel semantics only arises in somewhat exotic cases where teleosemantics would ascribe contents to states that are generated at random (Shea 2007b, pp. 427–30).

The advantage of varitel semantics is more substantial when we come to explanation. My accounts show why contents are suited to explaining how systems perform their functions—they do so by making use of exploitable relations carried by their components. Teleosemantics is more in the business of

typing representations by the behaviour they produce than in the business of explaining how a system produces behaviour.[14]

8.5 Ambit of Varitel Semantics

(a) Representation only if content is explanatory?

Is content in the eye of the beholder? Does the question of whether a system has representational properties depend upon whether it is useful for some observer to treat it as such? Not according to my accounts of content. Each account of content that falls within the varitel framework is given in terms of correlation, structural correspondence, robust outcomes, stabilization, and so forth. Why is that complex property of interest? Because it arises in the world for natural reasons and, where it arises, it generally **(p.211)** affords distinctive kinds of generalization and induction. Saying these properties allow us to engage in a distinctive form of explanation does not imply that the existence of these properties is relative to the existence of an observer, able and willing to go in for explanations of this kind. The existence of these patterns is an observer-independent fact, as is the fact that some properties explain others (§4.2b).

Since contents are not constituted by the explanatory practices of an observer, there is no requirement that contents should afford useful explanations in every case. Contents constructed in the ways I have set out are suited to getting explanatory purchase in many cases, but may not do so across the board. Consider for example a thermostat, one that is only slightly more sophisticated than the philosopher’s standard example. At input, it has two ways of gauging room temperature, a sensor for levels of sunlight and a sensor for thermal expansion. At output it controls the temperature through operating a radiator valve and an external air vent. Its capacity to keep a room at a set temperature is slightly better than a normal single input thermostat. It predicts the warming effect of sunlight and so smooths out some of the bumps in temperature that a standard thermostat would produce. But it is only slightly better. Arguably, holding the room temperature constant is a robust outcome produced by thermostat. Deliberate design may constitute that outcome as a task function. And it is a task function achieved by (simple) internal workings that make use of exploitable relations between internal states and distal states of the environment (there is some bridging).

However, this is a case where we would get little or no additional explanatory benefit from the representational explanation than we get from a non-semantic causal explanation of how proximal inputs generate proximal outputs, and how that affects the temperature of the room. It is a case where there is representational content but not one where representational explanations are substantially better than non-representational or vehicle-based explanations of behaviour.

Part of the reason why representational contents vary in their explanatory purchase is that the features which give them explanatory bite come in degrees. A system can produce outputs more or less robustly: across wider or narrower ranges of proximal inputs, and by more or fewer different proximal bodily movements. Furthermore, task functions can arise from a range of different stabilizing processes: natural selection, feedback-based learning, and simple contribution to persistence of a biological organism. In paradigm cases all three stabilizing processes will be at work and will be pulling in the same direction. In less paradigmatic cases they may come apart, so there can be different contents constituted relative to different task functions. In marginal cases, or in thought experiments like swampman, maybe only one of these processes comes into play; for example, contribution to persistence on its own. Such cases will have representational content if they meet one of the conditions above. However, in these more peripheral cases representational content may well have less explanatory purchase than in paradigm cases.

Task functions constituted only by design, as in the thermostat example above, may also give rise to contents that are less explanatorily useful, depending on the amount of **(p.212)** robustness and internal vehicle-based complexity that has been built into the system. On the other hand, some designed artefacts like a sophisticated computer-guided missile, with a high degree of internal complexity and the capacity to produce outputs very robustly, may be cases where the representational description is practically indispensable for explaining behaviour, in roughly the way Dennett claimed beliefs and desires are indispensable in practice for explaining human behaviour.

In short, representational contents are not in the eye of the beholder. They are constituted by the coming together of the properties mentioned in the conditions set out in previous chapters. The extent to which the representational explanations they afford are useful or practically indispensable will vary with the facts of the case.

Finally, I briefly assess whether the accounts of content put forward here are pragmatist. This label is sometimes associated with the claim that the brain exists to guide action, that it is therefore not in the business of building models of the world (Barrett 2011), and that we should not therefore expect representation to play a central role in the cognitive sciences (Anderson and Chemero 2016). Varitel semantics largely agrees with—indeed is based on—a version of the first claim. But it strongly rejects the second and third. The brain forms representations, and builds models of the world, and does so in the service of guiding action. Representational content in our case studies depends ineliminably on action, its functional significance, and the role of representations in guiding action. So varitel semantics has contents that are pragmatist in the sense that their content is action-derived, while rejecting the anti-representationalist conclusion that is drawn by those who claim that enactive

engagement with the world displaces representational content (Hutto and Satne 2015).

The regularity with which I use the word ‘explanation’ suggests pragmatism of another kind; my emphasis on the observer-independence of content properties suggests otherwise. Simon Blackburn offers a useful characterization of what pragmatists are often up to (Blackburn 2010). The pragmatist aims to explain why we go in for a certain kind of discourse. I am certainly doing that. A major explanandum for my project is the pattern of representational explanation found in the cognitive sciences. I can be seen as offering a (realist) explanation of why psychologists and cognitive neuroscientists engage in that kind of discourse and think about organisms and systems in that way.

What divides my approach from pragmatism is the kind of explanation I offer. An explanation is pragmatist, according to Blackburn, when it eschews any use of the referring expressions of the discourse but proceeds by talking in different terms about what is done by the discourse; for example, by showing that it serves a particular function (Blackburn 2010, pp. 1–2). By contrast I argue that, in order to explain behaviour in the kinds of cases I am interested in, we do have to refer to representations (as real vehicles) and their contents. What I have offered is a meta-semantic metaphysical account of what various theoretical terms in use in cognitive science refer to—terms in the family of representation, semantic information, content, correctness condition, satisfaction condition, and so on. In explaining how this discourse works I use the **(p.213)** terms representation and content as referring expressions, so my accounts of content are not pragmatist by Blackburn’s lights.

(b) Are any cases excluded?

Does varitel semantics imply that every natural system is processing internal representations? Won’t any system designed by natural selection, learning, or human design end up having internal states that count as representations? We have already seen that magnetotactic bacteria do not have representations (§8.2b). Similarly, the ‘two component’ signalling processes that are ubiquitous in bacteria (Lyon 2017) will be excluded when they depend on detecting just a single proximal cue.[15] Varitel semantics does extend readily to non-psychological cases, but there are principled reasons why many cases are excluded.

Consider the way plant roots follow local concentration gradients so as to move towards water (Takahashi 1997). Is the root (or the plant) representing the direction of water? No, because here the story about stabilization does not involve distal facts, but just how the root reacts to the proximal availability of water. Or consider a germinating seed that uses gravity to grow upwards towards the surface of the soil. The adaptively relevant fact here is distal, the availability of sunlight. Must it not have an internal representation of the direction of the sun? Must a plant that rotates in the direction of solar radiation

have an internal representation of the direction of the sun? These input–output mappings, as described, are not the basis of task functions because they are not robust outcome functions. The output is modulated just by intrinsic properties of the sensory input. Distal features of the environment are adaptively important, but there is no task function (stabilized function + robust outcome function) that concerns distal features of the environment, and no mechanism that 'bridges' across multiple sensory inputs in correlating with a distal feature of the environment.[16] The representational content that would be attributed by standard teleosemantic accounts marches exactly in step with a factorized explanation of the plant's behaviour. These are not representing systems according to varitel semantics.

Which is not to say that plants never go in for representation, let alone that representational content is restricted to psychological systems. Consider a plant that opens its flowers in the day and closes them at night. Suppose it just relies on changes in temperature, which alter internal biochemical processes. The opening and closing behaviour is produced in response to only one input, and so would not be a task function. It would be an evolutionary function of the plant, but would lack the robustness to be a task function. Now supplement the case slightly, making it more biologically realistic, so that the plant is also sensitive to light levels, giving it a second way of detecting **(p.214)** that evening has arrived. Then the plant has two ways of detecting that it is evening, and so the flower-closing behaviour would be a (very simple) task function of the plant. Internal processes in the plant could then well be representations with descriptive content about the time of day and with directive content telling it when to open and close its flowers.

But surely robustness is not unusual, but an absolutely ubiquitous feature of biological organisms? Cells have robust metabolic networks (Krasensky and Jonak 2012). Cells are even able to explore and sample new possible metabolic networks when they are put under conditions of severe stress (Szalay et al. 2007). During cell development the spindle microtubules that structure the cell grow in the right places robustly—they do so because of a process of selection in which many spindles are started and only those that reach their targets are preserved (Kirschner and Gerhart 1998, pp. 8422–3). Both metabolism and cell development produce outcomes robustly, and do so as a result of stabilizing processes that operate on both the phylogenetic and the ontogenetic timescale. Surely these are paradigm cases of robust outcome functions? Not as I have been using the term. Our target is cases where the same outcome is produced in response to different external inputs and is robust in the face of changes in the distal circumstances in which it is produced. These cellular and metabolic examples are cases of robust internal processes. They do show adaptive responses to things happening at the cell surface, such as damage to the cell wall, but the functions there are described in terms of intrinsic properties of the cell and changes happening to it. These are not cases of the kinds of functions

that can ground representational content, which is what my definition of task functions was designed to capture.

There is no reason why hormonal signals should be excluded from being representations within the varitel framework. Sometimes their operation will be explicable in purely intrinsic functional terms, but in many cases hormones are involved in tracking adaptively relevant distal facts by multiple routes (e.g. about conspecifics), part of the basis of task functions. Another set of cases is found in the immune system. It has complex mechanisms for detecting threats and responding adaptively, so it is very likely to help the organism perform task functions, and it would not be at all surprising if representations were involved in carrying out those task functions. In short, cases of internal subpersonal representation in organisms extend well beyond the psychological. Furthermore, these are cases where representation gets explanatory purchase: world-involving contents allow us to see how the hormonal system or immune system enables the organism to achieve certain distal outcomes in its environment.

Another way that subpersonal representations may arise internally is when there are task functions of systems that are smaller than the whole organism. For example, an individual cell is likely to have task functions, as may a larger unit like the immune signalling system. That system may have robust outcome functions concerning facts about other parts of the body distal to it. For example, an individual cell may have multiple ways of detecting the overall physiological state of the organism it is in (e.g. stressed vs. non-stressed) and responding appropriately. If so, the cell could have **(p.215)** a task function, one that could be representationally mediated. By this route some processes going on within an organism might count as task functions, functions for systems that are smaller than the organism as a whole. Caution is needed here, though. Not all evolutionary functions will count. To be a task function, an output of a system needs to have been the target of a stabilizing process operating on that system as such. Learning-type processes within a cell count. The process of generate-and-test used by the immune system may also count. But outputs of an internal system that are stabilized only because of the way they contribute to a stabilizing process operating at the level of the whole organism do not thereby count as task functions of the internal system.

Hormonal signalling may be like that, its functions deriving from task functions of the organism in which it operates. It may not be sufficiently distinct to count as a system in its own right; and if it does, there may be no stabilizing process acting at the level of the hormonal system as such, independently of those operating at the level of the whole organism. The same may apply to the brain. There are lots of processes of selective retention going on in the brain, of course, but these play a role in whole-organism stabilizing processes (various

kinds of learning), which counts against them having an intra-organism task function of their own.

Varitel semantics is more demanding than theories of content that are based just on evolutionary functions or just on correlational information. But that is not because it is designed to capture only a class of cases which seem, a priori, to count as calling for representations. So, it is no kind of desideratum that only psychological systems should be capable of forming representations. Instead, we took psychological systems as studied by cognitive science as a paradigm and aimed to account for what gives representational contents their explanatory purchase in those cases. We then find that that distinctive explanatory scheme extends more widely (§6.5b). For example, it includes the kinds of internal signalling in plants we've just discussed. It also covers many cases of between-organism animal signalling. Both the vervet's alarm call and the honeybee's nectar dance are cases where the representation producer integrates multiple cues in order to produce the signal. So, output behaviour is a task function, both in the monkeys and the honeybees.

When we come to the most familiar examples of representations in psychological systems, namely human conscious states and human beliefs and desires, it is very likely that there are things that set them apart from representational contents in non-psychological cases. They have further features relevant to content determination (§8.9), that are missing in non-psychological cases like animal signalling, plant tropisms, and hormonal signalling. My pluralism allows that the right account of content determination applicable to these cases might well be different. The special features of consciousness, or of the practice of offering and assessing reasons for action at the personal level, for example, might well make an important difference to content determination. If so, there will be an account of personal-level content (or more than **(p.216)** one) which applies only to psychological cases. However, I would resist the urge to identify these as the real or true accounts of content, since they would not apply to many other, subpersonal, psychological cases, where representational contents do have a clear explanatory role to play. Accounts of content that are plausibly restricted to the psychological are too narrow, and we have seen that accounts of content adequate to capturing the way representational explanations work in subpersonal psychology do indeed extend to some non-psychological cases.

8.6 Development and Content

Various theorists writing about concepts have noted that there is a tight connection between the circumstances in which a concept develops and the object or property it refers to. With representations less sophisticated than concepts a similar connection is often apparent. In previous work on artificial neural networks I explored the way that vehicles of content themselves develop as a result of training, that is as a result of reacting to samples in the environment and being tuned based on feedback about performance (Shea

2007a). This developmental process leads clusters to form in hidden layer state space. Those clusters are vehicles of content. They represent properties of the samples that caused their development.

If content is fixed by synchronic properties of a system, for example its causal sensitivity, then it is far from obvious why there should be any connection between the circumstances in which a representational vehicle develops and what it represents. Fodor was so puzzled by this phenomenon that he gave it a name, the DOORKNOB-*doorknob* problem (Fodor 2008). For Fodor the problem arises because it seems that many concepts are neither innate (i.e. unlearned), nor constructed out of existing concepts. (The problem is discussed in Shea 2016.) The concept DOORKNOB is neither present at birth, nor is it plausibly constructed out of other concepts like ROUND, ATTACHED TO A DOOR, FOR TURNING, etc. Instead, it is acquired by a thinker as a result of experience, a particular kind of experience: interaction with … doorknobs. Fodor was puzzled as to why that should be; puzzlement that arises in part, I suggest, from implicitly rejecting the idea that the circumstances in which the concept develops could play a role in fixing its content. After all, for Fodor a knock on the head could fortuitously put a thinker in a new brain state such that they have the concept DOORKNOB.

There are many empirically studied cases where the causes of the development of a new representational resource figure in its content. An example is recognizing new people by their faces. We acquire the ability to recognize an individual by seeing and perhaps interacting with that individual for a short time. The new recognitional capacity that results is probably dependent on a neural representation in the fusiform face area of the brain (Kanwisher 2000, Cohen and Tong 2001). The person who caused the development of this new recognitional capacity, which is mediated by a new vehicle of content, ends up being its referent. This is very similar to my neural network example **(p.217)** (Shea 2007a), where new state space clusters represent properties of the samples that caused their development. Game-theoretic sender–receiver models also show how new representations arise as a result of stabilizing processes (Skyrms 2010); for example, the replicator dynamics can lead senders and receivers to coordinate on a way of categorizing a range of stimuli (O’Connor 2014).

Laurence and Margolis (2002) have an account of the acquisition of natural kind concepts that links their content closely to the circumstances of their development. The child develops a new natural kind concept as a result of seeing a member of the kind:

> She sees a new object that has features that suggest that it is a natural object of some sort … upon encountering the item, the child releases a new

> mental representation and begins accumulating information about the object and linking this to the representation.
>
> [2002, p. 42]

So, an object which falls under a new kind causes the child to acquire a representation which refers to the kind, a representation which is used to store information about the new kind. Laurence and Margolis picture a process in which an existing contentless mental symbol is taken off the shelf and put into the right mind–world relations to constitute it as a concept of the kind. In my neural network case, interaction with samples in the world causes the development of a new vehicle which has the appropriate mind–world connections to have a certain content. Vehicle development and content development occur in tandem, due to the same causal process. In both cases there is a tight connection between the circumstances of development and the content of the new representations that result.

My accounts of content make that conclusion entirely unsurprising. Contents are fixed relative to task functions. Task functions arise as a result of some stabilizing process. Learning is a key case. I argued that outcomes that are the target of stabilizing processes are often stabilized and robustly produced as a result of internal mechanisms, mechanisms that make use of exploitable relations between internal components and the world. One common way that can happen is when the stabilizing process—for example, learning—gives rise to the internal mechanism which is responsible for an outcome being produced robustly and stabilized. Since contents are fixed by reference to outcomes that were stabilized and conditions that obtained which explain why those outcomes were stabilized, it is entirely unsurprising that contents should often concern properties of the objects the system was interacting with during development of a new representation (i.e. during the process of stabilization, that being the process by which content is constituted).

So, a feature which proponents of a synchronic metaphysics of content need to explain, and which poses a puzzle for Fodor, turns out to be readily explicable in the varitel framework. When a new representation develops as a result of interactions between a system and its environment, it will often end up representing the objects and properties causally involved in its development.

(p.218) 8.7 Miscellaneous Qualifications

In this section I go through a miscellaneous series of clarifications and qualifications.

In §2.3, ‘Externalist Explanandum, Externalist Explanans’, I argued that we should expect contents to be fixed by some complex relational property of representational vehicles. That would make contents suited to explaining how a system achieves distal outcomes in its environment. I have just argued that my

accounts fulfil that promise (§8.2). A system will have representational contents based on relational properties that bridge across a variety of different inputs and outputs, and thus are part of world-involving real patterns. Does it follow that contents can only concern distal features of the organism's environment and distal outcomes?

The answer is no. Proximal inputs like sensory properties, and proximal outputs like bodily movements, can be represented. UE information can concern internal states of the system. The point about explanatory purchase requires that the system should have some task functions that concern distal outcomes. It follows that it should have some descriptive representations about distal features of its environment. But that does not imply that every representation must concern something distal. An organism may also represent proximal inputs and outputs, as a means for calculating what is the case and what to do. For example, it may represent possible motor programs and use them in calculating which motor program needs to be executed in order to achieve some directively represented distal outcome in the current circumstances. Or it may keep track of sensory properties as a means for learning how to behave in new circumstances. Whether proximal features are represented in addition to distal ones depends on what is called for by the algorithm the organism is using in order to achieve its distal task functions.[17] Task functions are world-involving, and some representations in a system must be too, but it is not a requirement on being UE information that the correlation should concern a condition distal to the system (nor for UE structural correspondence).

A clear example of that is meta-representation. Some computations call for representations which represent the content of other representations. That arises in the relatively low-level system that does model-free reinforcement learning, since the algorithm involves a stage where the reward expected for an action is compared to the reward actually received and the difference is used to update reward expectations for the future (Shea 2014c). Varitel semantics can readily accommodate, both representations whose content concerns the content of other internal states, and representations that concern non-contentful internal states of the system (e.g. sensory states, bodily states and other internal properties).

A second caveat: I need to qualify the way I have discussed outputs produced by an organism. I have talked as if all outputs are bodily movements or the consequences of **(p.219)** bodily movements. In fact other kinds of output can also qualify; for example, releasing a chemical, producing an electrical discharge, or changing colour. Although movements have taken centre stage in all our discussions, everything I have said should be taken to cover outputs in general (when the other conditions for content are met, e.g. the output is or leads to an outcome which is task function of the organism).

Another significant oversimplification is found in the way I have talked about vehicles of content as constituent parts of an organism or other system. That is, I have pictured them as proper parts, or components of a mechanism. Those are the easiest cases to understand, and this conception covers all of our case studies, however the account is not restricted to such cases. Vehicles need not be proper parts of the system. Syntactic types can depend on properties of the whole system.

I don't know of any actual cases, but for illustration imagine a cell that is simultaneously undergoing three cyclical physiological processes, each something like the Krebs cycle, but involving the whole cell rather than a series of constituents. Take a dynamic systems approach to the cell. Each cyclical process can be occurring in a range of ways, which I'll call states of the cycle. Cycle C can be in various cyclical states C_1, or C_2, and so on; cycle D in states D_1, D_2, etc. Cycle C undergoes changes between cyclical states in a way that is affected by the states of cycles D and E, and vice versa. The whole system exhibits attractors, and perhaps bifurcations, and is affected by states of the environment. Dynamic properties of the whole cell, like being in state C_1 and state D_2, could in principle be vehicle properties, carrying world-involving contents, and interacting in ways that, by obeying generalizations about dynamic interactions amongst the C_i, D_i and E_i, are faithful to their contents. So, vehicles need not be proper parts of the system doing the representing. Furthermore, they need not form a mechanism (assuming not every causal interaction calls for a mechanism).

Next, a brief note on how my approach relates to evolutionary game theoretic models of signalling, communication, and meaning. These models were developed by Brian Skyrms and others (e.g. Skyrms 2010), following David Lewis's decision-theoretic treatment of signalling games (Lewis 1969). For Skyrms, the meaning of a signal in a signalling game is just a matter of the correlational information it carries, in particular how much it changes the probability that world states obtain or actions will be performed.

Shea et al. (2017) argue that these models need to be supplemented with a richer conception of meaning in order to account for phenomena like misrepresentation and deception. These phenomena arise in discussions of the models but are not given a formal treatment. We call this kind of meaning 'functional content', contrasted with the purely informational content put forward by Skyrms. In our treatment, functional content only arises at an equilibrium. It would be possible to apply our definition to non-equilibrium states. Functional content is essentially a matter of how signals are involved in generating rewards in certain world states, given the way receivers act on the signals. Signals are involved with generating rewards whether or not the population is an equilibrium. However, Shea et al. (2017) focus on a kind of content that only arises in an equilibrium state. Varitel semantics is broadly

similar. Content depends on **(p.220)** task functions. Task functions must be stabilized functions, which means they must have contributed to a stabilizing process like natural selection, learning, or contribution to an organism's persistence (leaving aside task functions based on deliberate design). So, they must, in some broad sense, have been attractors of the dynamic interactions between system and environment. The system need not currently be in an attractor state, but it must have been in one for its states to have content.

Tying functional content to stabilizing processes might seem problematic in the light of recent work showing that, in finite populations, signalling can arise in non-equilibrium states (Wagner 2012, 2015). For example, Wagner (2015) analyses a signalling game where populations converge on an attractor that is not a Nash equilibrium. Senders send signals that are informative about the world state and receivers act on them accordingly. My answer is simple. A stabilized function in my sense need not be a Nash equilibrium. The states Wagner identifies, being attractors to which the model converges, can thereby generate stabilized functions of the sender-receiver system. There may be a legitimate role for defining a notion of functional content for game theoretic models which applies to all states of the game, attractors and transients. However, my framework is motivated by the need to explain successful and unsuccessful behaviour. The parallel in game theoretic models is content that arises from attractor states. So, the restriction to defining functional content only in cases where there is or has been a stabilizing process (attractor) is suited to our purposes.

Moving on, there are some important issues about the nature of content, which should certainly be central to an account of conceptual content, that I have overlooked entirely. One is whether there is a level of content at the neo-Fregean level of sense; for example, a mode of presentation. I have been working only with referential content. Referential content, plus facts about vehicles—for example, that a system has two different vehicles concerned with representing colour—have together been adequate to explain all our target phenomena. Nevertheless, I want to remain neutral on whether a second level of content is justified in our simple cases, or indeed whether it turns out not to be needed even when we come to beliefs, desires, and concepts.

I am also setting aside questions of indexicality. For example, I assumed that particular locations figure in the content of the rat's spatial representations, but have not said whether they are picked out indexically or by non-indexical singular terms (§6.2d). Similarly, in the analogue magnitude case, the monkey's choice between two buckets of objects is correct if it picks the more numerous collection, but I have not said whether the analogue magnitude register for each bucket has the indexical content *that bucket contains n objects*, or whether the collection is picked out non-indexically as in *bucket A contains n objects* (or indeed whether the content is indeterminate between these possibilities). There

may be representations which the organism itself reuses in a variety of different contexts, where the right account says that content is a character, that is, a kind of content that combines with a context to deliver a truth condition. I remain neutral about how these issues should be dealt with.

(p.221) 8.8 How to Find Out What Is Represented

My accounts of content say what makes it the case that a simple system has representations with a certain content. It is concerned with the metaphysics of content, not with how we should find out what is represented. Nevertheless, varitel semantics has some straightforward implications for the epistemology of content.

The varitel framework is an elaboration of a procedure that is often used to establish content in cognitive science. Look at patterns of behaviour that are purposeful or adaptive and consider how the organism could perform in those ways by keeping track of aspects of the environment and calculating what to do when. That is, consider what algorithms could be producing the observed behaviour. Then search for evidence about internal workings in order to decide which possible algorithm is actual. Evidence of internal processes may be direct, through imaging, recording, or intervening in the brain; or indirect, through observing patterns of error, interference, and facilitation, like priming effects. When an algorithm which would produce the observed behaviour plausibly maps onto internal workings in the organism, then those elements are good candidates for vehicles of content, and the algorithm tells you what they represent. According to this picture, an early step is to look for robust outcome functions and assess whether they are also stabilized functions, and thus amount to task functions (i.e. outputs that are susceptible to representational explanation).

Considering task functions to be the target of representational explanation is seldom very explicit in cognitive scientific practice, however it is often implicit, regulating which kinds of behaviour are taken to be interesting and in need of representational explanation. More obvious is the search for information. Cognitive neuroscience directs a lot of energy at measuring the correlational information that is carried by different individual neurons, distributed arrays of neurons, and neural areas. My approach implies that not all information is relevant. Only information that is potentially germane to the task should be of interest. That restriction is often implicit in the scientific practice. Single unit recording investigates selectivity in respect of natural features of the distal world like lines, edges, and surfaces. Imaging usually looks for selectivity in respect of tasks or task-related features like faces, locations, object categories, and so on. So, in practice scientists are often in fact only interested in information that is potentially relevant to explaining how the organism behaves.

It is also implicit that information, to be relevant, has to be carried in a way that can be detected by downstream processing. When asking whether some neural area uses a rate code or a phase code, for example, a key consideration is whether the putative code can be read out by downstream processes. Katz et al. (2016) undermined the widely assumed importance of a signal in the lateral intraparietal area (LIP) by showing that knocking it out pharmacologically made no difference to behaviour. Hunt et al. (2012) formulate this requirement explicitly, noting that the information that can be decoded by an observer recording with an imaging technique or electrode can be quite different from 'the functional representations in the network, those used **(p.222)** by the brain' (p. 474). This makes explicit a constraint which is usually at work implicitly in cognitive neuroscientific practice.

The varitel framework does recommend some tweaks to current practice. When measuring correlational information, studies focus heavily on input sensitivity. Correlations with actions and outcomes are not completely overlooked, but they deserve greater emphasis, since output correlations almost always play a role in content determination. Furthermore, there could be a more explicit focus on the way behaviours are stabilized so as to underpin stabilized functions. In reward-based learning experiments, input correlations with reward delivered are always considered, output correlations with reward generated less so, although those output correlations are equally relevant. Indeed, it may be possible to generate quantitative measures, not just of correlational information, but of the way representational vehicles are involved with generating variable amounts of reward, along the lines of the reward-involving functional content vector defined by Shea et al. (2017).

Another mainstream way that content is investigated sits very naturally with varitel semantics. Investigators look for illusions and other systematic patterns of error. One might think that an error can only be identified once representational content has been ascertained, but in many cases problems at the level of behaviour, like vacillation, delay, or doing something clearly maladaptive, can be identified before being traced back to errors in what is being represented. My framework shows how we can be more rigorous about success and failure of behaviour. We need to consider the stabilization processes that have been at work; that is, what the organism has evolved and learned to do, and how its behaviour contributes to its persistence. These provide the standards against which success and failure of behaviour and its consequences should be judged. It follows that appeals to ethology and comparative psychology are of more than background interest. They throw light on evolutionary functions, which play a central role in constituting content.

8.9 Differences at the Personal Level

The book has not attempted to tackle personal-level contents. We have focused instead on the hopefully simpler question of how content arises in subpersonal representational systems. In this final section we look briefly at how various features of the personal level might make a difference to content determination.

First off, consciousness. The phenomenal character of conscious mental states may be fixed by intrinsic properties of the subject. More controversially, consciousness may in turn determine the representational content of those states. A naturalistic theory of consciousness seems a distant prospect. If so, we may be a long way from a theory of content for conscious states. On the other hand, there is some hope that the representational content of conscious states may determine their phenomenal character, in which case a theory of content for conscious states will be a route to a theory of consciousness. To follow that route, we need a better understanding of the distinctive **(p.223)** functional role of consciousness in order to see which aspects of the way conscious states operate may play a content-determining role. Relevant functional features could include: a global workspace, a drive for consistency between information from different modalities and subsystems, integration of descriptive information with valence and motivation, practical grasp of the enabling conditions for forming a representation reliably, a role in practical reasoning and learning for the future, storage in episodic and semantic memory, and feelings of confidence. Any or all of these features may play a role in content determination for conscious states. None is obviously reducible to the ingredients we have been working with so far.

A second potentially relevant feature is meta-representation or metacognition. On some views a thinker's object-level conscious state of seeing a red rose is simultaneously a meta-level state with a content along the lines of *I am currently seeing a red rose*. There could also be non-conscious mental states that have meta-representation built in. In either case, the fact that object-level and meta-level content are fixed in parallel would form an important part of the theory of content determination.

Thirdly, many theorists of concepts have thought that relations amongst concepts play an important role in determining content. A concept may encode information about how entities and properties are related causally and hierarchically, for example. This gives rise to deductive and inductive entailment relations between concepts. Alternatively, information about relations between categories may be encoded in a thinker's disposition to draw inferences between concepts. Furthermore, some ranges of properties are mutually exclusive, and some objects physically exclude one another in space. The relations of entailment and exclusion instantiated in a network of beliefs, or of concepts, may play a role in content determination in a way that has not been covered by the framework we have been using.

There may be an important difference between implicit and explicit connections between representations.[18] Suppose that when I see *that is a brown furry object of xyz shape/size* I am thereby disposed to think *that is a dog*.[19] This disposition implicitly encodes the information that objects that are brown, furry and of xyz shape/size tend to be dogs. That implicit representation is true. It's being true partly explains why I successfully behave in a dog-appropriate way on this occasion. My DOG concept also figures in some of my explicit representations; that is, beliefs about dogs. I believe that is okay to leave a docile dog alone with a young child. Suppose that belief is false. That has potentially disastrous consequences once I become responsible for young children. But I can modify my explicit belief by reasoning and reflecting on it. Becoming a parent, I start being aware of reports of seemingly docile dogs attacking when left alone with young children. So, I change my belief. Information represented implicitly in my dispositions to apply my DOG concept can also change as a result of experience. I can retrain **(p.224)** my inferential dispositions. But that is a different process from the way conscious deliberation changes my explicit beliefs. Both kinds of connection may be important in the story about content-determination for concepts. In particular, the special functional role of conscious deliberation in forming and changing explicit beliefs may have a special role in content-determination.

The meaning of beliefs and desires may also depend on interpersonal norms, and/or on the meaning of words, which may in turn depend on social processes (§6.5b). For content determination we would then have to cast the net more widely than a single individual, so as to include culturally based stabilizing processes like the patterns of transmission and use of a word in a social group.

As well as beliefs and desires that figure in episodes of thinking, people also have standing beliefs and desires. I have long believed that Lima is the capital of Peru, even though it is many months since I entertained that thought occurrently (until just now). There may be vehicles of the standing beliefs, stored away in long term semantic memory in the same way as data is stored on a computer disk. Or ascriptionism may be the best account (§1.3); for example, Dennett's intentional stance (Dennett 1981; see also Williams 2016, Williams 2018). Either way, standing beliefs may have observer-relative contents. On ascriptionist views there is no straightforward connection between the contents of standing beliefs and the contents of occurrent beliefs that are tokened in episodes of thinking. The content of standing beliefs could be observer-relative while occurrent states of thinking have non-observer relative contents (perhaps fixed in part by social processes, in the way just suggested). These are important content-relevant features of standing beliefs.

Even if additional functional features of the personal level play a content-determining role, should we nevertheless expect the overall varitel framework to apply, perhaps with an augmented menu of exploitable relations? Or, more

minimally, should we expect that the explanatory purchase of representational content will still depend on a convergence between task functions and internal workings which form an algorithm for their achievement? The answer is I don't know. It may do. But the richer features found at the personal level may underpin a different kind of account of content constitution. For example, if consciousness, fixed by intrinsic properties, fixes the content of conscious states, then content determination there works quite differently.

A natural thought here, but one that should be resisted, is the idea that subpersonal contents are picked up and used by personal-level processes. Subpersonal vehicles are constituted as contentful because of their relations: relations to features of the distal environment, to outputs of an algorithm implemented in the organism, and to stabilizing processes that have operated on the organism. Personal-level processes may capitalize on some of those same relations, for example the correlational information carried by a concept may be important to fixing its content. But that is not to make use of the content of a subpersonal representation. The contents of subpersonal representations are not things that are sitting around ready to be used by personal-level processes. It is a mistake to think of the content of a vehicle as a property that is routinely portable, **(p.225)** a property that would automatically be carried around if that vehicle is deployed in a different cognitive process. The absence of a straightforward connection here means there is no simple way that personal-level contents are determined by contents of any subpersonal representations they make use of. On the other hand, it frees our theorizing about subpersonal content from the need to play a role in accounting for personal-level content.

So I am remaining open-minded about what kinds of insight, if any, varitel semantics will offer into the nature of personal-level content. Might pluralism, at least, come in handy at the personal level? Not just a pluralism which allows that content at the personal level will be different, but a pluralism that expects different kinds of personal-level state to have their content determined differently? That, too, is open to debate. Close connections—for example, between belief content and the content of conscious states—may make it inappropriate to leave pluralism open about content determination as between occurrent beliefs and conscious states.

While I think it's too soon to venture an opinion about how the personal-level story will go, I would argue that the varitel accounts of content represent a substantial advance. We started the book with the question 'what is a thought process?', and with the worry that the powerful answer offered by the representational theory of mind would be undermined if we were unable to answer the related question 'what is representational content?' Now we have an answer to the content question that works for large swathes of the cognitive

sciences. Our optimism that we can answer the same question for representation at the personal level should therefore increase.

And it's not just general optimism that naturalism is taking us in the right direction. The varitel accounts of content give us a staging post, a fixed point from which to build. Intentionality is less mysterious now that we can see how cases of it arise through the coming together of some relatively well-understood natural properties. Psychology and cognitive neuroscience have excavated the computational processes that underlie some quite complex patterns of human behaviour. The foundational worry that those theories are based on a false assumption about meaning can now be assuaged. We can see correlation, correspondence, and function at work in these cases, giving rise to representational content in a completely un-mysterious way. So, we are now in a position to ask what needs to be handled differently to deal with personal-level cases. We have a reasonably detailed understanding of how personal-level representations are different, in ways that are relevant to content, as even the brief discussion above indicates. So, we have a good list of resources to draw on. Rather than being stuck at an impasse, with the lurking suspicion that the question is intractable, or representationalism entirely misconceived, we are now faced with a workable research programme—substantial and challenging, but with a clear sense of how to make progress.

That is a valuable payoff; however, the most important achievement of varitel semantics, if it succeeds, is to elucidate the nature of content in subpersonal cases. Subpersonal representation is a big challenge in its own right. The manifest success of **(p.226)** the cognitive sciences has seen representational theories deployed ever more widely. It is now pressing to understand the intentionality at the heart of these theories. I have argued that varitel semantics allows us to understand how those explanatory practices work. Huffing and puffing with information, function and structural correspondence can do the job. We can indeed give a naturalistic account of content in the brain and other subpersonal representational systems.

Notes:

([1]) I.e. at the same level of aggregation: semantic properties are properties of the very same objects (i.e. vehicles) as are found in a vehicle-based causal description of a system's operation.

([2]) The non-semantic, vehicle-based description says how inputs to the system effect changes to internal vehicles, how those in turn influence further vehicles, and how that internal process eventuates in bodily movements. It is simplest to think of this as a complete causal description, saying how the system would react to any kind of influence on it. However, the syntactic description is itself a set of special science generalizations, and as such there will usually be things that can happen to the system that it overlooks, that appear as exceptions to its

generalizations. For example, a stronger gravitational field might modulate the way a system performs actions, without that change being mediated by any differences in the vehicles involved in internal processing. A different kind of example is where an unusual influence on internal vehicles—e.g. through transcranial magnetic stimulation (TMS)—makes changes to an intermediate step of internal processing in a way that bypasses changes to upstream vehicles and so shows up as an exception to the causal transitions described by the algorithm.

([3]) This is not a novel feature of varitel semantics; e.g. a requirement for robust tracking (Sterelny 1995) or constancy mechanisms (Burge 2010) has the same effect.

([4]) The question of the observer-dependence or observer-independence of these patterns is orthogonal to another feature of Dennett's view, the fact that it does not commit to there being vehicles of content ('ascriptionism', §1.3). RTM's commitment to vehicles could be combined with the view that the contents represented by those vehicles are observer-dependent.

([5]) To be part of a real pattern, a content property underpinned by the bridged relation needs to show up in generalizations connecting it to others (e.g. in psychological theories). I am not committed to Dennett's particular way of theorizing real patterns in terms of Kolmogorov complexity. I am just relying on the idea that the generalizations are based on the underlying system mostly being organized in a particular way, and on that being a non-observer-dependent feature of the world (contra for example perspectivalism; Craver 2013). Many real patterns do compress information in the sense that they only allow a 'probabilistic recovery of the underlying system' (Ladyman 2017, p. 153).

([6]) It also probably lacks sufficient internal structure to count as implementing an algorithm by which it achieves its functions. Indeed, dead bacteria continue to rotate into alignment with a magnetic field (Cummins et al. 2006, Schulte 2015).

([7]) E.g. Hutto and Satne (2015) with respect to some forms of intentionality; Egan (2014) with respect to cognitive content.

([8]) That is not of course to deny that non-semantic properties of vehicles of content can be causally relevant. Causal relevance: an apple on the greengrocer's scales causes the spring to extend. Here the apple's mass is causally relevant and its colour is not.

([9]) Dretske says 'indicate', which is a more restrictive species of correlational information. Godfrey-Smith (1992) objects that natural selection does not

require indication but often makes use of weaker kinds of correlational information about adaptively relevant facts.

([10]) Those who point to the relevance of externalist properties to explaining the behaviour of an organism in its environment—e.g. to sorting behaviour (Davies 1991) or action guidance (Peacocke 1993)—are also arguably in the causal relevance camp, although theorists who object to a causal role for externalist properties will see these as being cases where content has merely heuristic value.

([11]) Dretske argues that this form of explanation of behaviour is unavailable when it is natural selection in ancestors, rather than learning, which explains why the organism acts on R to produce M as output (Dretske 1988, p. 94; 1991, pp. 206–7). My framework covers both cases. Godfrey-Smith (1992, pp. 294–6) argues that natural selection should be assimilated to Dretske's scheme of explanation.

([12]) But recall that 'computational structure' is not a case of structural representation: §5.7a.

([13]) It is agreed on all sides that representations produce correlations on the output side at the time of selection/stabilization: correlations with the distal effects they produce. This is a form of exploitable relation. However, that is not enough for there to be content according to my accounts. Exploitable relations at input also have to be in place: see §4.2a.

([14]) Godfrey-Smith (1996, pp. 171–95) argues that teleosemantics makes representational explanation of behaviour akin to explaining why a sleeping pill put a person to sleep by citing its dormitive virtue. In Shea (2007b) I argue that, while dormitive virtue explanations are not empty, adding an exploitable relation requirement—in that case a correlation requirement—makes the representational explanation more substantial.

([15]) Even if the mechanism is very complex, as described in Hsieh and Wanner (2010).

([16]) It is the absence of a distal-involving task function that stops representation arising. It is not a requirement on being UE information that the correlation should concern something distal (§8.7). (However, because of the requirement for a distal-involving task function, some items of UE information in the system need to concern distal features of the environment.)

([17]) There is a rough parallel here with Burge's view. He argues that the capacity to represent properties like time, for which there is no constancy detection mechanism, is derivative from the capacity to represent properties for which there is a constancy detection mechanism (Burge 2010).

([18]) The contrast between implicit and explicit representation I am using here is spelt out more carefully in Shea (2015).

([19]) *xyz shape/size* is some set of shape and size properties represented in visual experience.