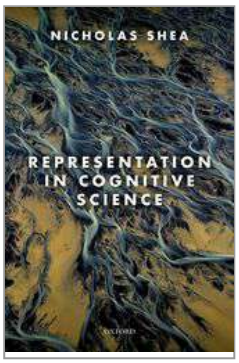

Representation in Cognitive Science
Nicholas Shea

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

Introduction

Nicholas Shea

DOI:10.1093/oso/9780198812883.003.0001

Abstract and Keywords

This chapter offers a breezy introduction to the content question, the question of what determines the content of a mental representation. Existing approaches are outlined: informational semantics, inferential role semantics, correspondence theories, ascriptionism and the intentional stance, and teleosemantics. This discussion highlights the major issues that the book's positive account must address if it is to succeed.

Keywords: intentionality, content question, informational semantics, inferential role semantics, correspondence theories, intentional stance, teleosemantics

1.1 A Foundational Question

The mind holds many mysteries. Thinking used to be one of them. Staring idly out of the window, a chain of thought runs through my mind. Concentrating hard to solve a problem, I reason my way through a series of ideas until I find an answer (if I'm lucky). Having thoughts running through our minds is one of the most obvious aspects of the lived human experience. It seems central to the way we behave, especially in the cases we care most about. But what are thoughts and what is this process we call thinking? That was once as mysterious as the movement of the heavens or the nature of life itself.

New technology can fundamentally change our understanding of what is possible and what mysterious. For Descartes, mechanical automata were a revelation. These fairground curiosities moved in ways that looked animate, uncannily like the movements of animals and even people. A capacity that had previously been linked inextricably to a fundamental life force, or to the soul, could now be seen as purely mechanical. Descartes famously argued that this could only go so far. Mechanism would not explain consciousness, nor the capacity for free will. Nor, he thought, could mechanism explain linguistic competence. It was inconceivable that a machine could produce different arrangements of words so as to give an appropriately grammatical answer to questions asked of it.[1] Consciousness and free will remain baffling. But another machine has made what was inconceivable to Descartes an everyday reality to us.

Computers produce appropriately arranged strings of words—Google even annoyingly finishes half-typed sentences—in ways that at least respect the meaning of the words they churn out. Until quite recently a ‘computer’ was a person who did calculations. Now we know that calculations can be done mechanically. Babbage, **(p.4)** Lovelace, and others in the nineteenth century saw the possibility of general-purpose mechanical computation, but it wasn’t until the valve-based, then transistor-based computers of the twentieth century that it became apparent just how powerful this idea was.[2]

This remarkable insight can also answer our question about thinking: the answer is that thinking is the processing of mental representations. We’re familiar with words and symbols as representations, from marks made on a wet clay tablet to texts appearing on the latest electronic tablet: they are items with meaning.[3] A written sentence is a representation that takes the form of ink marks on paper: ‘roses are red’. It also has meaning—it is about flowers and their colour. Mental representations are similar: I believe that today is Tuesday, see that there is an apple in the bowl, hope that the sun will come out, and think about an exciting mountain climb. These thoughts are all mental representations. The core is the same as with words and symbols. Mental representations are physical things with meaning. A train of thought is a series of mental representations. That is the so-called ‘representational theory of mind’.

I say the representational theory of mind is ‘an’ answer to our question about thinking, not ‘the’ answer, because not everyone agrees it is a good idea to appeal to mental representations. Granted, doing physical manipulations on things that have meaning is a great idea. We count on our fingers to add up. We manipulate symbols on the page to arrive at a mathematical proof. The physical stuff being manipulated can take many forms. Babbage’s difference engine uses gears and cogs to do long multiplication (see Figure 1.1). And now our amazingly powerful computers can do this kind of thing at inhuman speed on an astonishing scale. They manipulate voltage levels not fingers and can do a lot

more than work out how many eggs will be left after breakfast. But they too work by performing physical manipulations on representations. The only trouble with carrying this over to the case of thinking is that we're not really sure how mental representations get their meaning.

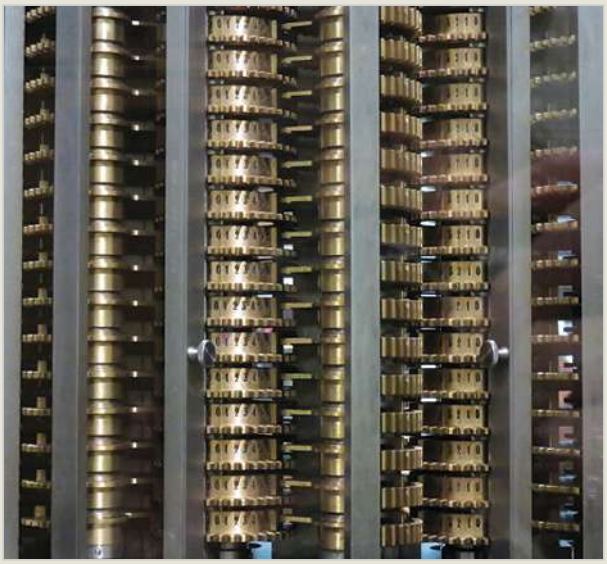

Figure 1.1 Babbage's difference engine uses cogs and gears to perform physical manipulations on representations of numbers. It is used to multiply large numbers together. The components are representations of numbers and the physical manipulations make sense in the light of those contents—they multiply the numbers (using the method of differences).

For myself, I do think that the idea of mental representation is the answer to the mystery of thinking. There is very good reason to believe that thinking is the processing of meaningful physical entities, mental representations. That insight is one of the most important discoveries of the twentieth century—it may turn out to be *the* most important. But I have to admit that the question of meaning is a little problem in the foundations. We've done well on the 'processing' bit but we're still a bit iffy about the 'meaningful' bit. We know what processing of physical particulars is, and how processing can *respect* the meaning of symbols. For example, we can make a machine whose manipulations obey logical rules and so preserve truth. But we don't yet have a clear idea of how representations could *get* meanings, when the meaning does not derive from the understanding of an external interpreter.

(p.5) So, the question remains: how do mental states[4] manage to be about things in the external world? That mental representations are about things in the world, although utterly commonplace, is deeply puzzling. How do they get their *aboutness*? The physical and biological sciences offer no model of how naturalistically respectable properties could be like that. This is an undoubted lacuna in our understanding, a void hidden away in the foundations of the cognitive sciences. We behave in ways that are suited to our environment. We do so by representing the world and processing those representations in rational ways—at least, there is strong evidence that we do in very many cases. Mental representations represent objects and properties in the world: the **(p.6)** shape of a fruit, the movement of an animal, the expression on a face. I work out how

much pasta to cook by thinking about how many people there will be for dinner and how much each they will eat. ‘Content’ is a useful shorthand for the objects, properties and conditions that a representation refers to or is about. So, the content of one of my thoughts about dinner is: *each person needs 150g of pasta*.

What then is the link between a mental representation and its content? The content of a representation must depend somehow on the way it is produced in response to input, the way it interacts with other representations, and the behaviour that results. How do those processes link a mental representation with the external objects and properties it refers to? How does the thought in my head connect up with quantities of pasta? In short: what determines the content of a mental representation? That is the ‘content question’. Surprisingly, there is no agreed answer.

This little foundational worry hasn’t stopped the cognitive sciences getting on and using the idea of mental representation to great effect. Representational explanation is the central resource of scientific psychology. Many kinds of behaviour have been convincingly explained in terms of the internal algorithms or heuristics by which they are generated. Ever since the ‘cognitive revolution’ gave the behavioural sciences the idea of mental representation, one phenomenon after another has succumbed to representational explanation, from the trajectories of the limbs when reaching to pick up an object, to parsing the grammar of a sentence. The recent successes of cognitive neuroscience depend on the same insight, while also telling us how representations are realized in the brain, a kind of understanding until recently thought to be fanciful. Figure 1.2 shows a typical example. The details of this experiment need not detain us for now (detailed case studies come in Part II). Just focus on the explanatory scheme. There is a set of interconnected brain areas, plus a computation being performed by those brain areas (sketched in the lower half of panel (a)). Together that tells us how participants in the experiment manage to perform their task (inset). So, although we lack a theory of it, there is little reason to doubt the existence of representational content. We’re in the position of the academic in the cartoon musing, ‘Well it works in practice, Bob, but I’m not sure it’s really gonna work in theory.’

The lack of an answer to the content question does arouse suspicion that mental representation is a dubious concept. Some want to eliminate the notion of representational content from our theorizing entirely, perhaps replacing it with a purely neural account of behavioural mechanisms. If that were right, it would radically revise our conception of ourselves as reason-guided agents since reasons are mental contents. That conception runs deep in the humanities and social sciences, not to mention ordinary life. But even neuroscientists should want to hold onto the idea of representation, because their explanations would be seriously impoverished without it. Even when the causes of behaviour can be picked out in neural terms, our understanding of why *that* pattern of neural activity produces *this* kind of behaviour depends crucially on neural activity being about things in the organism's environment. Figure 1.2 doesn't just show neural areas, but also how the activity of those areas should be understood as **(p.7)** representing things about the stimuli presented to the people doing a task. The content of a neural representation makes an explanatory connection with distal features of the agent's environment, features that the agent reacts to and then acts on.

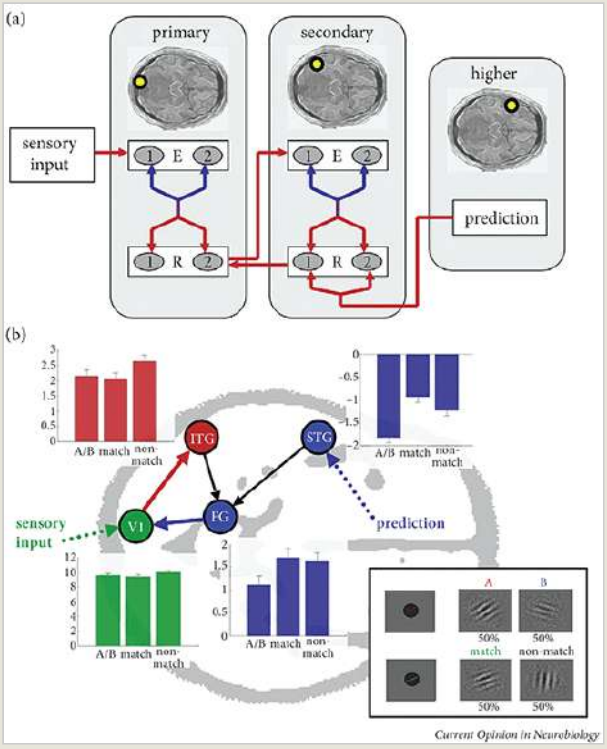

Figure 1.2 A figure that illustrates the explanatory scheme typical of cognitive neuroscience (from Rushworth et al. 2009). There is a computation (sketched in the lower half of panel (a)), implemented in some interacting brain areas, so as to perform a behavioural task (inset). The details are not important for present purposes.

One aspect of the problem is consciousness. I want to set that aside. Consciousness raises a host of additional difficulties. Furthermore, there are cases of thinking and **(p.8)** reasoning, or something very like it, that go on in the absence of consciousness. Just to walk along the street, your eyes are taking in information and your mind is tracking the direction of the path and the movement of people around you. Information is being processed to work out how to adjust your gait instant by instant, so that you stay on your feet and avoid the

inconvenience of colliding with the person in front engrossed in their smartphone. I say those processes going on in you are a kind of reasoning, or are like familiar thought processes, because they too proceed through a series of states, states about the world, in working out how to act. They involve processing representations in ways that respect their contents. Getting to grips with the content of non-conscious representations is enough of a challenge in its own right.[5]

The content question is widely recognized as one of the deepest and most significant problems in the philosophy of mind, a central question about the mind's place in nature. It is not just of interest to philosophers, however. Its resolution is also important for the cognitive sciences. Many disputes in psychology concern which properties are being represented in a particular case. Does the mirror neuron subsystem represent other agents' goals or merely action patterns (Gallese et al. 1996)? Does the brain code for scalar quantities or probability distributions (Pouget et al. 2003)? Do infants represent other agents' belief states or are they just keeping track of behaviour (Apperly and Butterfill 2009)? Often such disputes go beyond disagreements about what causal sensitivities and behavioural dispositions the organism has. Theorists disagree about what is being represented in the light of those facts. What researchers lack is a soundly based theory of content which tells us what is being represented, given established facts about what an organism or other system responds to and how it behaves.

This chapter offers a breezy introduction to the content question for non-experts. I gesture at existing literature to help demarcate the problem, but proper arguments will come later (Parts II and III). So that I can move quickly to presenting the positive account (Chapter 2 onwards), this chapter is more presupposition than demonstration. It should bring the problem of mental content into view for those unfamiliar with it, but it only offers my own particular take on the problem.

1.2 Homing In on the Problem

The problem of mental content in its modern incarnation goes back to Franz Brentano in the nineteenth century. Brentano identified aboutness or 'intentionality'[6] as being a peculiar feature of thoughts (Brentano 1874/1995). Thoughts can be about objects and properties that are not present to the thinker (the apple in my rucksack), are distant in time and space (a mountain in Tibet), are hypothetical or may only lie far in the future (the explosion of the sun), or are entirely imaginary (Harry Potter). How can mental **(p.9)** states reach out and be about such things? Indeed, how do beliefs and perceptual states manage to be about an object that is right in front of the thinker (the pen on my desk), when the object is out there, and the representations are inside the thinker?

We could ask the same question about the intentionality of words and natural language sentences: how do they get their meaning? An obvious answer is: from the thoughts of the language users.[7] The meaning of a word is plausibly dependent on what people usually take it to mean: 'cat' is about cats because the word makes people *think* about cats. That kind of story cannot then be told about mental representations, on pain of regress. In order to start somewhere we start with the idea that at least some mental representations have underived intentionality. If we can't make sense of underived intentionality somewhere in the picture—of where meaning ultimately comes from—then the whole framework of explaining behaviour in terms of what people perceive and think is resting on questionable foundations. The most fruitful idea in the cognitive sciences, the idea of mental representation, which we thought we understood, would turn out to be deeply mysterious, as difficult as free will or as consciousness itself.

When asked about the content of a familiar mental representation like a concept, one common reaction is to talk about other mental states it is associated with. Why is my concept DOG about *dogs*?[8] Because it brings to mind images of dogs, the sound of dogs barking, the feel of their fur and their distinctive doggy smell. We'll come back to these kinds of theories of content in the next section, but for now I want to point out that this answer also just pushes back the question: where do mental images get their contents? In virtue of what do they represent the visual features, sounds, tactile properties and odours that they do? Underived intentionality must enter the picture somewhere.

The task then is to give an account of how at least some mental representations have contents that do not derive from the contents of other representations. What we are after is an account of what determines the content of a mental representation, determination in the metaphysical sense (what makes it the case that a representation has the content it does?) not the epistemic sense (how can we tell what the content of mental representation is?). An object-level semantic theory gives the content of mental representations in a domain (e.g. tells us that cognitive maps refer to spatial locations). Many information-processing accounts of behaviour offer a semantic theory in this sense. They assign correctness conditions and satisfaction conditions to a series of mental representations and go on to say how those representations are involved in generating intelligent behaviour. Our question is a meta-level question **(p.10)** about these theories: in virtue of what do those representations have those contents (if indeed they do)? For example, in virtue of what are cognitive maps about locations in the world? Our task then is to formulate a meta-semantic theory of mental representation.

It is popular to distinguish the question of what makes a state a representation from the question of what determines its content (Ramsey 2007). I don't make that distinction. To understand representational content, we need an answer to both questions. Accordingly, the accounts I put forward say what makes it the

case, both that some state is a representation, and that it is a representation with a certain content.

We advert to the content of representations in order to explain behaviour. To explain how you swerve to avoid an oncoming smartphone zombie, the psychologist points to mental processes that are tracking the trajectory of people around you. A theory of content can illuminate how representations play this kind of explanatory role. One central explanatory practice is to use correct representations to explain successful behaviour. That assumption is more obviously at work in its corollary: misrepresentation explains failure. Because she misperceived the ground, she stumbled. Because he thought it was not yet eight o'clock, he missed the train. Misrepresentation explains why behaviour fails to go smoothly or fails to meet an agent's needs or goals. When things go badly for an agent, we can often pin the blame on an incorrect representation. We can also often explain the way they do behave; for instance, misrepresenting the time by fifteen minutes explains why he arrived on the platform nearly fifteen minutes after the train left.

Misrepresentation is one of the most puzzling aspects of representational content. A mental representation is an internal physical particular. It could be a complex pattern of neural activity. Cells firing in the hippocampus tell the rat where it is in space so it can work out how to get to some food at another location. If the cell firing misrepresents its current location, the rat will set off in the wrong direction and fail to get to the food. Whether a representation is correct or incorrect depends on factors outside the organism, which seem to make no difference to how the representation is processed within the organism (e.g. to how activity of some neurons causes activity of others). Yet its truth or falsity, correctness or incorrectness, is supposed to make a crucial explanatory difference. The capacity to misrepresent, then, is clearly a key part of what makes representational content a special kind of property, a target of philosophical interest. Any good theory of content must be able to account for misrepresentation.

A theory of content need not faithfully recapitulate the contents relied on in psychological or everyday explanations of behaviour. It may be revisionary in some respects, sometimes implying that what is actually represented is different than previously thought. Indeed, a theory of content can, as I suggested, help arbitrate disputes between different proposed content assignments.[9] However, it should deliver reasonably determinate contents. A theory of content needs to be applicable in concrete cases. **(p.11)** For example, reinforcement learning based on the dopamine subsystem explains the behaviour elicited in a wide range of psychological experiments. We can predict what people will choose if we know how they have been rewarded for past choices. Plugging in facts about what is going on in a concrete case, a theory of content should output correctness conditions and/or satisfaction conditions for the representations involved. The

determinacy of those conditions needs to be commensurate with the way correct and incorrect representation explains successful and unsuccessful behaviour in the case in question. A theory of content would obviously be hopeless if it implied that every state in a system represents every object and property the system interacts with. Delivering appropriately determinate contents is an adequacy condition on theories of content.

The problem of determinacy has several more specific incarnations. One asks about causal chains leading up to a representation. When I see a dog and think about it, is my thought about the distal object or about the pattern of light on my retina? More pointedly, can a theory of content distinguish between these, so that it implies that some mental representations have distal contents, while others represent more proximal matters of fact? A second problem is that the objects we think about exemplify a whole host of properties at once: the dog is a member of the kind *dog*, is brown and furry, is a medium-sized pliable physical object, and so on. The *qua* problem asks which of these properties is represented. Finally, for any candidate contents, we can ask about their disjunction. A theory may not select between *that is a dog* and *that is a brown, furry physical object* but instead imply that a state represents *that is a dog or that is a brown, furry physical object*. Rather than misrepresenting an odd-looking fox as a dog, I would end up correctly representing it as a brown furry object. If every condition in which this representation happens to be produced were included, encompassing things like shaggy sheep seen from an odd angle in poor light, then the representation would never end up being false. Every condition would be found somewhere in the long disjunction. We would lose the capacity to misrepresent. For that reason, the adequacy condition that a theory of content should imply or explain the capacity for misrepresentation is sometimes called the ‘disjunction problem’. The *qua* problem, the disjunction problem, and the problem of proximal/distal contents are all different guises of the overall problem of determinacy.

Since we are puzzled about how there could be representational contents, an account of content should show how content arises out of something we find less mysterious. An account in terms of the phenomenal character of conscious experience, to take one example, would fail in this respect.[10] Standardly, naturalistic approaches offer accounts of content that are non-semantic, non-mental, and non-normative. I am aiming for an account that is naturalistic in that sense. Of course, it may turn out that there **(p.12)** is no such account to be had. But in the absence of a compelling a priori argument that no naturalistic account of mental representation is possible, the tenability of the naturalistic approach can only properly be judged by the success or failure of the attempt. The project of naturalizing content must be judged by its fruits.

1.3 Existing Approaches

This section looks briefly at existing approaches to content determination. I won't attempt to make a case against these approaches. Those arguments have already been widely canvassed. My aim is to introduce the main obstacles these theories have faced, since these are the issues we will have to grapple with when assessing the accounts of content I put forward in the rest of the book. Although the theories below were advanced to account for the content of beliefs, desires, and conscious states, the same issues arise when they are applied to neural representations and the other cases from cognitive science which form the focus of this book.

One obvious starting point is information in the weak sense of correlation.[11] Correlational information arises whenever the states of items correlate, so that item X's being in one state (smoke is coming from the windows) raises the probability that item Y is in another state (there is a fire in the house). A certain pattern of neural firing raises the probability that there is a vertical edge in the centre of the visual field. If the pattern of firing is a neural representation, then its content may depend on the fact that this pattern of activity makes it likely that there is a vertical edge in front of the person.

Information theory has given us a rich understanding of the properties of information in this correlational sense (Cover and Thomas 2006). However, for reasons that have been widely discussed, representational content is clearly not the same thing as correlational information. The 'information' of information-processing psychology is a matter of correctness conditions or satisfaction conditions, something richer than the correlational information of information theory. Sophisticated treatments use the tools of information theory to construct a theory of content which respects this distinction (Usher 2001, Eliasmith 2013). However, the underlying liberality of correlational information continues to make life difficult. Any representation carries correlational information about very many conditions at once, so correlation does not on its own deliver determinate contents. Some correlations may be quite weak, and it is not at all plausible that the content of a representation is the thing it correlates with most strongly.[12] A weak correlation that only slightly raises the **(p.13)** chance that there is a predator nearby may be relied on for that information when the outcome is a matter of life or death. Representations often concern distal facts, like the presence of a certain foodstuff, even though they correlate more strongly with a proximate sensory signal. Furthermore, a disjunction of conditions is always more likely than conditions taken individually: for example, an object might be an eagle, but it is more likely that it is an eagle *or* a crow. So, content-as-probability-raising faces a particularly acute form of the disjunction problem. Correlational information may well be an ingredient in a theory of content (Chapter 4), but even the sophisticated tools of mathematical information theory are not enough, without other ingredients, to capture the

core explanatory difference between correct representation and misrepresentation.

Another tactic looks to relations between representations to fix content. We saw the idea earlier that the concept DOG gets its meaning from the inferences by which it is formed, such as from perceiving a brown furry object, and perhaps also from conclusions it generates, such as inferring that this thing might bite me (Block 1986). Patterns of inferences are plausibly what changes when a child acquires a new mathematical concept (Carey 2009). They are also the focus of recent Bayesian models of causal learning (Gopnik and Wellman 2012, Danks 2014). Moving beyond beliefs to neural representations, dispositions to make inferences—that is, to transition between representations—could fix content here too. If all inferences are relevant, holism threatens (Fodor and Lepore 1992): any change anywhere in the thinker's total representational scheme would change the content of all of their representations. There have been attempts to identify, for various concepts, a privileged set of dispositions that are constitutive of content (e.g. Peacocke 1992). However, it has proven difficult to identify sets of inferences that can do the job: that are necessary for possessing the concept, plausibly shared by most users of the concept, and sufficiently detailed to be individuative—that is, to distinguish between different concepts. For these reasons, inferential role semantics has not had much success in naturalizing content, except perhaps for the logical constants. The same concerns carry over when we move from beliefs to other representations relied on in the cognitive sciences.[13]

Relations amongst representations may be important for another reason. They endow a system of representations with a structure, which may mirror a structure in the world. For example, spatial relations between symbols on a map mirror spatial relations between places on the ground; and that seems to be important to the way maps represent. In the same way, Paul Churchland has argued that the similarity structure on a set of mental representations of human faces determines that they pick out certain **(p.14)** individuals (Churchland 1998, 2012). Taken on its own the correspondence idea produces an implausibly liberal theory of representation (Cummins 1989, Godfrey-Smith 1994a, Shea 2013c). As we will see, however, structural correspondence is another plausible ingredient in a theory of content (Chapter 5).

Another group of theories are ascriptionist: they ascribe mental states to the whole person based on how she behaves, but don't commit to mental representations being physical particulars. This eschews what I described above as the core insight of the representational theory of mind (RTM). I discuss ascriptionism here because it remains a viable alternative to RTM (Williams 2016, 2018), especially for beliefs and desires,[14] so it will be important to be clear about the explanatory advantages that flow from the commitment to representations as physical particulars, when that commitment is warranted

(see §2.5). Donald Davidson’s version of the view derives from rational decision theory (Davidson 1974a, 1974b). The choices of an agent who obeys some minimal conditions of rationality can be modelled as if the agent has an ordered set of preferences about states of the world combined with a set of probabilistic beliefs about the likelihood of world states and the chance that actions she can perform would bring about other world states. According to Davidson, for an agent to be interpretable in this way is a key part of what it is to have beliefs and desires, to be a representer.

Daniel Dennett’s intentional stance is in the same family of views (Dennett 1981). He emphasizes that there is nothing irrealist about this approach. People and other agents are tied up in patterns of interaction with the world that we can predict and explain from the intentional stance—that is, by treating them as having beliefs and desires. We could not do so if these interactions were described in purely physical terms, for example in terms of energy transduced by sensory receptors, producing neural states, which generate movements of the limbs. I can arrange to meet a colleague at a café in far-away Canberra three months hence. The intentional stance allows me to predict where they will be at 10 a.m. on 1 July in a way that would be impossible in practice by just keeping track of their moment-by-moment physical interactions with their environment. Even if it were not intractable, although a purely physical description would tell us, in physical terms, what is going to happen instant-by-instant, it would miss out on real patterns that exist in the behaviour (Dennett 1991). Those real patterns only come into view when we take the intentional stance, but the patterns are there irrespective of whether we recognize them (see §2.3 and §8.2b). The ontology of patterns means there is an observer-independent fact of the matter about which systems are interpretable from the intentional stance.

Dennett’s ascriptionism is a tenable and substantial naturalistic account of representational content.[15] In this sense we already have a good theory of content. It is realist **(p.15)** about what it takes to be a representer. However, I will reserve the term ‘realism’ for accounts that are committed to there being real *vehicles* of content: individuable physical particulars that bear contents and whose causal interactions explain behaviour. As I have been describing the problem of mental content, realism about vehicles is a core part of what it takes to be a mental representation. There are many cases where we have good evidence for realism about mental representations. We have already seen some examples in passing; subsequent chapters go into detail about many more. Where there are vehicles, representational explanation can explain more (§2.5). So, my task is to formulate an account of content applicable to cases where we have good reason to be realist about mental representations.

1.4 Teleosemantics

Teleosemantics is the final stop on our whistle-stop tour of the problems faced by existing theories of content. We will look at this family of views in slightly more detail since teleosemantics is the closest precursor to the accounts presented in this book. Teleosemantic views are those that give etiological functions a content-fixing role. This does not exclude there also being a role for correlational information, inferential role, or structural correspondence. A second commitment of the teleosemantic views of Ruth Millikan and David Papineau is a central role for a representation *consumer*: an identifiable subsystem that takes representations as input and generates outputs in response (Millikan 1984, 1989, Papineau 1987, 2016).

Peter Godfrey-Smith calls a commitment to representation consumers the ‘basic representationalist model’ (Godfrey-Smith 2006). It goes beyond the standard representational theory of mind (RTM), that is the commitment to representations as causally interacting particulars. The central idea of the basic representationalist model is that a representation is a stand-in that is relied on by a consumer to allow it to deal with some external state of affairs (see Figure 1.3). The consumer uses the state X as a guide to something else Y that it does not have access to directly. The idea is not that the consumer reads or interprets the representation, but simply that it reacts to **(p.16)** an intermediate state in a particular way. For example, ‘consumer’ honeybees observe dances of incoming bees as a guide to where nectar is located. In that case the representation is out in the open. In most psychological cases the representation is internal and the consumer is a subsystem within the organism. Ruth Millikan’s teleosemantic theory is also committed to there being an identifiable representation producer.

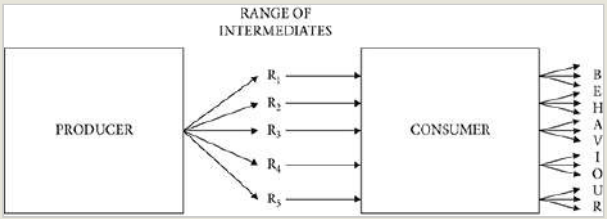

Figure 1.3 The basic representationalist model.

Informational approaches to content direct our attention to the way a representation is produced. Conditions in the world cause a representation to be tokened;[16] the representation having been produced raises the probability that those conditions obtain. Consumer-based views invert the picture. Downstream effects on a consumer both constitute states as representations and fix their contents. What a representation means depends on how it is used downstream, on what it is taken to mean by a consumer of the representation. If the organism is relying on R as a stand-in, the way the consumer behaves in response to R will betray what it is taking R to mean.[17] Papineau’s version of this idea targets beliefs and desires in the first instance (Papineau 1987, but see Papineau 2003). To see what a person believes, see how they act to satisfy their desires. So, the content of a belief is roughly the

condition under which the behaviour it prompts would satisfy the consumer's desires. Sitting at my laptop, an electronic sound prompts an internal state R in me, which causes me to click on an icon to open my inbox. Given my desire to read messages sent to me, state R has the content *there is a new email message for you*. The representation R is separate from the consumer subsystem. The content of the representation derives from the way the consumer reacts to R.

For Millikan, the content of a representation is the condition under which behaviour of the consumer, prompted by the representation, will be successful (Millikan 1984). The distinctive contribution of teleosemantics is to understand success in evolutionary terms. The behaviour of the consumer subsystem has evolutionary functions. Success is a matter of performing those evolutionary functions so as to promote survival and reproduction. The success conditions for a behaviour are the conditions which obtained when behaviour of that type was selected. They are conditions which explain why behaviour of that type led systematically to survival and reproduction.

Consider the way honeybees communicate about where to forage (Figure 1.4). Incoming bees that have found a source of nectar are producers. They perform a dance that indicates the location of the nectar. The direction of the dance correlates with the direction of nectar and the time spent waggling correlates with distance. Outgoing bees are consumers. They condition their foraging behaviour on the dance. The dance acts as a stand-in for the location of nectar, something the outgoing bees have no direct **(p.17)** access to. The behaviour of the consumer is to fly off in a direction and for a distance that corresponds to the dance they have observed and then to start foraging at that location. This pattern of behaviour is very likely to be the result of natural selection on ancestor colonies of bees. For each type of dance there is an associated specific condition; for example, two seconds of vertical waggling might correspond to there being nectar 400 metres away in the direction of the sun. That is the condition under which behaviour of consumers in the past prompted by dances of that form led systematically to survival and reproduction. There being nectar 400 metres away in the direction of the sun is part of a direct explanation of why behaviour of that type has been stabilized by natural selection. (Millikan also places considerable weight on there being a systematic relationship between different dances and their corresponding locations, which I discuss further in §5.5.)

(p.18) Millikan coined the term ‘Normal explanation’ for this kind of evolutionary explanation of how representation-prompted behaviour of the consumer was selected (Millikan 1984). What is evolutionarily Normal may be statistically rare, for example that a sperm actually fertilizes an ovum. The Normal cases are the ones that matter for natural selection. A complete Normal explanation would go into all kinds of details about the mechanism, and might also mention various background factors, like gravity remaining constant. Millikan’s focus is the least detailed Normal explanation of the specific type of behaviour prompted by a representation R. For the bee dance, this mentions the presence of nectar 400 metres from the hive, but not details of the implementational mechanism, nor the fact that gravity remained constant.

Figure 1.4 The dance of the honeybee indicates the location of a source of nectar.

Consumer behaviours will generally have a nested set of evolutionary functions: to fly off a certain distance and direction, to forage there, to obtain nectar, and to promote survival of the hive and production of offspring. Not all of these figure in the content-fixing story. It follows from what Millikan says about consumers making use of mapping rules that there will be different Normal explanations of different behaviours prompted by different representations. So, we can’t simply explain all the dances by the fact that there was nectar somewhere nearby when they were selected. Content is fixed relative to the behaviour of the consumer specific to each representational vehicle. That excludes general evolutionary functions of a behaviour like promoting survival of the hive. For the same reasons there is considerable specificity in the success condition associated with each type of behaviour: getting nectar 400 metres away rather than just getting nectar.

In short, teleosemantics founds content on evolutionary functions, and standard teleosemantics also depends upon there being a special kind of causal structure, a separation between representations and their consumers. Teleosemantics is the basis for a good account of content in some simple representational

systems,[18] of which animal signalling cases are a paradigm: macaque alarm calls, beaver tail slaps for danger, and the honeybee nectar dance.[19]

1.5 Challenges to Teleosemantics

Teleosemantics may be a good theory of content for animal signalling, and perhaps also for some kinds of internal communication that are the direct products of natural **(p.19)** selection, like hormonal signalling,[20] but there are significant obstacles to applying the theory more widely, to mental representations in general. The purpose of this section is not to demonstrate that teleosemantics fails but, as with the other theories I have mentioned, to say what its main challenges are thought to be, so that the accounts of content I put forward in later chapters can be assessed against those challenges.

The first obstacle to consumer-based teleosemantics is the need to identify a representation consumer with the right properties to play a content-constituting role (both to make it the case that some internal states are representations and to fix their contents). In real psychological cases behaviour depends on the interaction of several different representational vehicles. There need be no identifiable subsystem that takes this collection of representations as input and produces behaviour as output. Instead, the whole organism relies on interactions between internal representations in order to initiate and guide appropriate behaviour. We could instead take the outputs to be internal: representations rather than behaviour. Then each layer of processing could act as a consumer for representations lower down in the hierarchy. But it is not clear whether there is a non-circular account of evolutionary success conditions if the outputs that constitute content are further representations.

Nor does psychological processing always divide into neat layers to allow it to be parcelled up in this way (as we will see in Chapter 4). Some of the most compelling support for realism about representations (for RTM) comes from cases where something is known about the neural states that drive behaviour. Representation in the brain is however particularly unsuited to the consumer-based treatment, for the reason we have just seen: it is very hard to see a principled way to identify representation consumers in the brain, if consumers are devices whose output fixes content (Cao 2012). Even idealized diagrams of neural circuitry are notoriously interactive, consisting of a complex mix of feedforward, feedback, and lateral connections, some proceeding in a series of levels or layers, others cross-cutting or bypassing those layers (see Figure 1.5). That is reflected in information-processing accounts of how representations interact to drive behaviour. I don’t take this to be a knock-down argument against consumer-based views, but it will be an advantage of my account that I do not have to appeal to consumers to play a content-constituting role (Chapters 3–5).

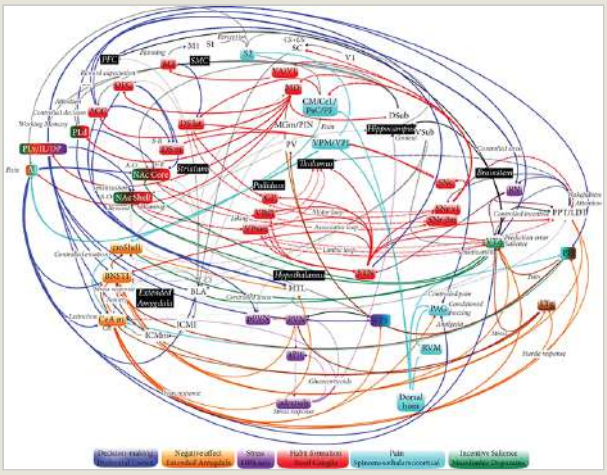

Figure 1.5 Some of the functional connections in the rat brain that are important for reward-guided behaviour (from George and Koob 2010).

The second challenge for teleosemantic theories of content is to formulate a notion of etiological function that is suited to playing a content-constituting role. Both Millikan and Papineau appeal to biological functions. Biological functions are based on evolution by natural selection. Most representation-types we are looking at are learnt. They have not evolved directly by natural selection. Millikan argues that new vehicles produced by learning have derived functions, functions that derive from the purpose of the **(p.20)** learning mechanism. For example, the mechanism in infants that tracks faces and learns patterns in visual input from faces plausibly has the evolutionary function of enabling the infant to reidentify individuals (more precisely, to behave in ways that depend on reidentifying the same individual again, e.g. its mother). That evolutionary function is relational. Which particular individuals it is supposed to track depends on who the baby has been interacting with. So, when the mechanism operates in baby Zeb and learns the pattern that is distinctive of his father Abe, the new representation has the derived function of tracking that particular individual, Abe.

This account works when the learning mechanism has a specific (relational) evolutionary function. But much learning in higher animals, especially in humans, is the result of general-purpose learning mechanisms. For example, the function of classical conditioning is just to find patterns in the sensory input. Such general evolutionary functions do not ground specific functions for acquired representations. Suppose we hear a loon's distinctive song. General-purpose learning mechanisms in the brain can track regularities in acoustic input. So, we learn the characteristic acoustic pattern of the song. On hearing part of the song, the mechanism can complete the pattern. Is this supposed to track loons in general, or an individual loon, or a distinctive pattern in the incoming sound waves or auditory neural input? The very general evolutionary function **(p.21)** of the learning mechanism does not decide between these. So relational evolutionary proper functions are not a promising basis for grounding all content.

Millikan argues that functions can also arise directly from learning when this involves a selection process in its own right (Millikan 1984, p. 45, see also Papineau 1987, pp. 65–7). She thinks this applies to instrumental conditioning. Dretske (1988) also puts forward a theory of content based on instrumental conditioning. An animal acts on a stimulus in a new way, which generates a reward, and so cements the disposition to act in that way. The reward is delivered because of some aspect of the stimulus; for example, the light indicated that there would be a peanut on the right-hand side, and the animal learnt to reach to the right-hand side in response to the light. The aspect of the correlational information carried by the stimulus which explains why this action tendency was stabilized constitutes the content of the new representation formed in the process.

Dretske's account is not based on the evolutionary function of instrumental conditioning. It is a basis of content determination, in a system that undergoes learning, that is not derivative from evolutionary functions at all. Nor does it depend on assimilating learning to a process of generate-and-test like natural selection (Kingsbury 2008). It suggests that we should have a broader conception of the kinds of stabilizing processes that can give rise to content. Indeed, Dretske's theory is part of the inspiration for my approach whereby there are several different stabilizing processes that can ground content (Chapter 3). However, Dretske's account of how content explains behaviour only applies to one kind of learning mechanism, instrumental conditioning (Dretske 1988, pp. 92–5; 1991, pp. 206–7). The question for teleosemantic theories of content is to specify which kinds of stabilizing processes give rise to the kind of etiological functions that ground content—and to explain why a certain way of delimiting etiological function is the right one for theories of representation to rely on.

The third main challenge faced by teleosemantics is highlighted by the swampman thought experiment. Swampman is an imaginary molecule-for-molecule duplicate of a person (created, let's say, by random chance when lightning hits a swamp). Teleosemantics implies that swampman has no representational states, because he has no evolutionary history. Some have taken intuitions about swampman to be a basis for objecting to a theory of content. As we will see shortly, intuitions have little probative value for our kind of project (§2.2). Nevertheless, the swampman case is important, because it highlights an implication of the theory. It forces us to reflect on whether there are good reasons for representational content to be based on history.

At first pass, representational explanation does not seem to depend on evolutionary history at all. By recognizing that behaviour was driven by a representation of the location of an object, say, it seems that the cognitive scientist is picking out a synchronic property of the organism. It also seems that the representational vehicle, for example a syntactic or neural state, is a

synchronic cause of the behaviour. How internal processing unfolds, and hence how the organism will make bodily movements, is caused moment-to-moment by the intrinsic properties of representational vehicles. It follows **(p.22)** that, if we take an organism that has evolved by natural selection and had a lifetime of interacting with its environment, make a duplicate with all the same internal properties, and place the duplicate in just the same environment, we will be able to make the same predictions about how it will behave.

Millikan argues that the intrinsic duplicate falls outside the real kind which underpins our inductive practices (Millikan 1996). Our inductions about people and their representational properties go right, to the extent that they do, because humans form a historical kind, sharing descent from a common ancestor, shaped by natural selection. That answer just pushes back the question however. It doesn't say why there are not also non-historical kinds that would enter into the same kinds of explanation. The fact that predictions would go right for swampman suggests that there is some synchronic property shared by humans, that would also be shared by their intrinsic ahistorical duplicates, and which underpins inductions.

The teleosemanticist should pause at this point and ask us to focus on the explanandum, the thing which representational contents are called on to explain. We point to representations to explain how organisms and other systems manage to interact with their environment in useful and intelligent ways. The explanandum is a pattern of successful behaviour of a system in its environment. That explanandum is absent at the moment swampman is created. It's not just that swampman has not yet performed any behaviour. (He already has dispositions to behave in certain ways.) It is that it's quite unclear that some behaviours should count as successful and others not. So, the creature with no history has no contents but that is fine because it has nothing which contents are called on to explain.

The 'no explanandum' argument does not rescue standard teleosemantics, however (see §6.4). It may show that we have no explanandum at the moment of swampman's creation, but it does not show that deep evolutionary history is needed for there to be an explanandum in place. As soon as an intrinsic duplicate starts interacting with its environment, stabilizing processes will begin to operate. It will do things that contribute to its persistence as an organism. It will undergo learning: behavioural patterns will be repeated or altered based on feedback. Doing things again that have been stabilized in the past looks like a kind of success—these are behaviours that have contributed to survival of the organism and the persistence of these behavioural dispositions (in the recent past). So, an organism's individual history seems to be enough to set up an explanandum which representational contents could be called on to explain.

Dretske's learning-based theory of content calls for individual learning history but not evolutionary history (Dretske 1988). It shows that there being something to explain—such as how the organism succeeds in getting food—does not depend on evolutionary history. So, an organism's learning history, taken on its own, seems to be enough to ground: (i) an explanandum, concerning the organism's interactions with its environment; and (ii) a kind of representational content suited to explaining those interactions. What thinking about swampman shows us is that teleosemantics lacks a **(p.23)** good account of why—given its explanatory role—representational content should be the kind of property that depends on *evolutionary* history.

Finally, I come to an objection that can be made to some teleosemantic accounts of content, and which also applies to varying extents to other naturalistic theories of content. How does content get its explanatory purchase? What does it add to a purely causal description of how a system operates and how it interacts with its environment to label some of its states with content? Dretske gave an answer to this question, arguing that contents figure in 'structuring cause' explanations, explaining why a system is wired the way it is, rather than synchronic causal explanations (Dretske 1988). That is an exception, however. Most theories of content, while telling us how content is determined, have relatively little to say about why content determined in that way has a special explanatory role (e.g. Fodor 1991). We turn to this issue in the next chapter, which sets out a framework for content-determination specifically designed to elucidate the explanatory role of content. We return to it again in Chapter 8 once we have detailed accounts of content in hand. **(p.24)**

Notes:

([1]) Descartes (1637/1988, p. 44: AT VI 56: CSM I I40), quoted by Stoljar (2001, pp. 405–6).

([2]) Developments in logic, notably by Frege, were of course an important intermediate step, on which Turing, von Neumann, and others built in designing computing machines.

([3]) We'll have to stretch the point for some of my son's texts.

([4]) I use 'mental' broadly to cover all aspects of an agent's psychology, including unconscious and/or low-level information processing; and 'state' loosely, so as to include dynamic states, i.e. events and processes. 'Mental state' is a convenient shorthand for entities of all kinds that are psychological and bear content.

([5]) Roughly, I'm setting aside beliefs and desires (doxastic states) and conscious states—see §2.1. I use 'subpersonal' as a label for mental representations that don't have these complicating features.

([6]) This is a technical term—it's not about intentions.

([7]) Another tenable view is that sentences have underived intentionality. For beliefs and desires, the claim that their content derives from the content of natural language sentences has to be taken seriously. But here I set aside the problem of belief/desire content (§2.1) to focus on simpler cases in cognitive science.

([8]) I use small caps to name concepts; and italics when giving the content of a representation (whether a subpropositional constituent, as here, or a full correctness condition or satisfaction condition). I also use italics when introducing a term by using (rather than mentioning) it.

([9]) E.g. whether infants are tracking others' mental states or just their behaviour.

([10]) That is not in itself an argument against such theories—it could turn out that intentionality can only be properly accounted for in phenomenal terms—but it is a motivation to see if a non-phenomenal theory can work.

([11]) Shannon (1948) developed a formal treatment of correlational information—as a theory of communication, rather than meaning—which forms the foundation of (mathematical) information theory. Dretske (1981) applied information theory to the problem of mental content.

([12]) Usually the strongest correlational information carried by a neural representation concerns other neural representations, its proximal causes, and its effects. The same point is made in the existing literature about beliefs. My belief that there is milk in the fridge strongly raises the probability that I have been thinking about food, only rather less strongly that there actually is milk in the fridge.

([13]) Concepts (the constituents of beliefs) are usually thought to have neo-Fregean sense, as well as referential content (content that contributes to truth conditions). We may well have to appeal to inferential relations between concepts to account for differences in sense between co-referential concepts, and/or to vehicle properties (Millikan 2000, Sainsbury and Tye 2007, Recanati 2012). This book does not deal with concepts. I leave aside the issue of whether we need to appeal to neo-Fregean senses in addition to vehicle properties and referential contents.

([14]) Neither Davidson nor Dennett claimed that their ascriptionism could be extended to the neural representations that are characteristic of the case studies we consider here.

([15]) He advances it not for neural representations but as an account of belief-desire content. Davidson's account is not naturalistic in our sense. He argues that it is not possible to give an account of content in non-normative terms.

([16]) A representation is *tokened* when an instance of it is realized. E.g. the rat has an array of place cells that represent locations. One of these representations is tokened when a place cell is active.

([17]) The same idea is in Braithwaite (1933): I believe that p means that, under relevant external circumstances, relative to my needs, I will behave in a manner appropriate to p. Braithwaite also anticipated a naturalistic treatment of what it is for an action to be appropriate to a person's needs: 'satisfaction of these needs is something of which I do not despair of a naturalistic explanation'. Success semantics has the same structure (Whyte 1990).

([18]) Even there, in my view standard teleosemantics needs to be supplemented with a further requirement, so that it is not just an output-oriented theory of content (Shea 2007b). The requirement is that a representation should carry correlational information about the condition it represents (more carefully: that the putatively representational state should have carried correlational information at the time selection was taking place).

([19]) Ethological work on animal signalling identifies exactly the same factors as relevant to the content of an animal signal: what the signal correlates with, the behaviour that is produced in response, the evolutionary function of that behaviour, and the conditions that matter for fulfilling that function (Searcy and Nowicki 2005, p. 3).

([20]) Also genetic information: it shows that genes carry semantic information, throws light on what genetic information can be called on to explain, and also applies to other forms of inheritance systems, i.e. of signalling between generations (Shea 2007b, 2009, 2011b, 2012a, 2012b, 2013a, Shea et al. 2011).