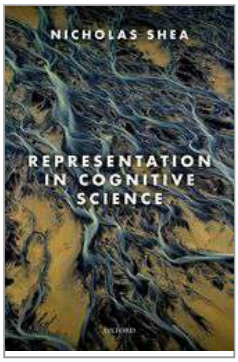

Representation in Cognitive Science
Nicholas Shea

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

Functions for Representation

Nicholas Shea

DOI:10.1093/oso/9780198812883.003.0003

Abstract and Keywords

What kind of functions are suited for grounding representational content? Do they derive from behaviour that is robust and apparently goal-directed, or from consequence etiology? Rather than choosing between these two elements the account here combines them: ‘robust outcome functions’ combine with ‘stabilized functions’ to form ‘task functions’, which are the functions-for-representation that offer a good basis for fixing content. Task functions allow space for contents on which they are based to have a distinctive kind of explanatory purchase.

Keywords: teleology, etiological function, stabilization, robust outcome, task function, representational explanation

3.1 Introduction 47
3.2 A Natural Cluster Underpins a Proprietary Explanatory Role 48
3.3 Robust Outcome Functions 52
3.4 Stabilized Functions: Three Types 56
(a) Consequence etiology in general, and natural selection 56
(b) Persistence of organisms 57
(c) Learning with feedback 59
(d) A ‘very modern history’ theory of function 62
3.5 Task Functions 64
3.6 How Task Functions Get Explanatory Purchase 67
(a) Illustrated with a toy system 67
(b) Swamp systems 69

3.7 Rival Accounts 72
3.8 Conclusion 74

3.1 Introduction

Varitel semantics has two variable elements: functions and exploitable relations. Chapters 4 and 5 look at exploitable relations. This chapter deals with functions. To apply our framework, we need to specify what it is for there to be a task being performed by an organism or other system. These tasks are the functions it is supposed (in some sense) to perform. Philosophical work on function has mostly focused on naturalizing biological functions, for which the constraints may be different. We are after a notion of function that is suited to figure in a theory of content: function-for-representation. Philosophical theories of function are often tested against intuitions about what counts as a function, a malfunction, a side effect, or an outcome with no kind of function at all. Intuitions can bear little weight for our purposes. Instead, our theorizing is guided by the goal of explaining representational explanations of behaviour (the desideratum in §2.2).

I take from teleosemantics the idea that natural selection is a source of functions that are partly constitutive of content. However, evolutionary function is too narrow **(p.48)** (§1.5). When behavioural dispositions are the result of a general-purpose learning mechanism, the evolutionary function of the learning mechanism does not deliver specific functions for the newly learnt behaviours. This chapter argues that behaviours of an individual organism can acquire functions as a result of its interaction with its environment, independently of which evolutionary functions it has. Furthermore, swampman suggests that we can representationally explain the behaviour of complex organisms, interacting with an environment, irrespective of their evolutionary history. Neither of these considerations is a decisive objection to teleosemantics' claim that the functions relevant to representation must ultimately find a basis in natural selection. However, they do motivate us to look for a way of specifying the task being performed by a system, for the purpose of varitel semantics, that does not depend on its deep evolutionary history.

My account of function will combine two elements. These broadly correspond to the two strands of Aristotelian teleology: a functional outcome is a natural occurrence that comes about always or for the most part, and that is for the sake of something (Shields 2013). The first corresponds to my robust outcomes: an organism is disposed to achieve the outcome in a range of circumstances and tends to pursue the outcome in the face of obstacles until it is achieved. The second strand is *consequence etiology*: an organism produces an outcome because of the consequences that flow from it. How can behaviour be caused because of its consequences? That in turn is naturalistically explicable if the outcome has been the target of some stabilizing process: the outcome is caused

now partly because of consequences of producing the same type of outcome in the past.

Rather than choosing between these two elements, as most philosophical theories of function do, my account combines them (§3.2). Robust outcome functions (§3.3) combine with stabilized functions (§3.4) to form task functions (§3.5), which are the functions-for-representation which I argue are a good basis for fixing content. Section 3.6 explains how task functions give representational content its explanatory purchase and do so in a way that need not depend on evolutionary history. Finally, §3.7 briefly compares some rival accounts of function in the literature.

3.2 A Natural Cluster Underpins a Proprietary Explanatory Role

Humans and other animals are paradigmatic representation-using systems. Animal behaviour achieves a range of outcomes robustly. Complex internal workings (engaging representations) are involved in doing so. Those outcomes often contribute to survival and/or reproduction. And a consequence etiology applies: an animal has the disposition to produce these outcomes partly because outcomes of the same type have been produced in the past, when they contributed to survival of the individual, or were the targets of learning, or of natural selection. That is, they have been the target of stabilizing processes. The clustering of a certain kind of causation by internal workings **(p.49)** with robustness and stabilization underpins the explanatory purchase of representational explanation.

The cluster exists for a reason. When robustness is not due to external constraints, the disposition to produce outcomes robustly is not usually possessed by accident. Often a stabilizing process has been responsible for the system acquiring robust outcome functions. An example that does not involve representations is sex determination. Since they produce an important outcome, mechanisms of sex determination have been the target of natural selection. A variety of backup mechanisms have evolved to ensure that the suite of traits needed to be a male, say, reliably come on stream together. Natural selection has made the outcome robust.

The most basic robustness tactic which evolution has hit on is survival itself. Survival of an individual organism is survival of its behavioural dispositions. Death of an organism is a form of non-robustness of all of its behavioural dispositions. It is no accident that producing outcomes robustly goes along with surviving, nor that robustly produced outcomes tend to contribute to the survival of the organism. One might object here that natural selection is really only about reproduction. Survival of the individual is at best subsidiary, and many traits are directed at reproduction in a way that compromises survival (Griffiths 2009). That is obviously correct: not all adaptations are survival-conducive. However, our project is not to define the ambit of natural selection, but to look for patterns

in nature. From that perspective, it is striking that so much behaviour in the animal kingdom is conducive to survival. That is because it has contributed to reproduction by contributing to survival. Because that way of being selected is so widespread, biologists typically conceive of natural selection in terms of contribution to reproduction *and* survival. Natural selection has given us a huge array of complex systems that maintain themselves in a state that is out of equilibrium with their environment (homeostasis) and act in ways that promote their own survival.

Evolution's other great robustness trick, exemplified in animal behaviour, is learning. Learning when a particular behaviour promotes survival is a way of making survival more robust. Learning new ways to behave generates new routes by which outcomes can be produced—general outcomes like survival and reproduction, and also more specific outcomes like avoiding a predator or getting a foodstuff. Learning new circumstances in which to perform, or new routes to generating, a behavioural outcome is a way of making that outcome more robust. Learning, like evolution, is a stabilizing process by which behavioural outcomes come to be robustly produced.

These three stabilizing processes—natural selection, learning, and contribution to survival—are at work throughout the animal kingdom. Each is a way that production of an outcome in the past contributes to raising the chance that an outcome of the same type is produced again. That is, each is a form of consequence etiology. They are all processes, on different timescales, which make production of an outcome of a particular type more likely. Furthermore, both learning and evolution are ways that a particular behaviour can come to be produced more robustly: learning from feedback allows an organism to overcome obstacles or learn new routes to producing an outcome; and **(p.50)** evolution can canalize a selected outcome so it is produced more robustly. Contributing to survival is not on its own a mechanism by which behaviours come to be produced more robustly, but for a biological organism, which is a complex out-of-equilibrium self-producing system (§3.4b below), producing outputs that contribute to its own persistence is an indispensable prerequisite for survival, which as we have seen is biology's most basic robustness trick. These are the reasons why robust outcomes tend to have been the target of one or more of these stabilizing processes. Robustness and stabilization come together in our cluster.

For example, one robust outcome function observed in the behaviour of mountain chickadees (*Poecile gambeli*) is their disposition to return to a previously cached item of food, doing so in a variety of conditions, from a variety of starting locations, and when the food is hidden in different ways (Pravosudov and Clayton 2001). Consider an individual chickadee, call her Jayla. Jayla's having retrieved cached food in the past is a cause of her survival. So, when she retrieves a food item now, obtaining cached food in the past is a contributory

cause. Obtaining food is such a basic need that it is also the target of several learning mechanisms. Jayla's having this form of behavioural disposition now is partly explained by the outcomes it has produced in Jayla's past, namely obtaining food. So, obtaining cached food is, on the basis of learning, a stabilized function of Jayla's behaviour. Furthermore, learning in this way has doubtless been the result of natural selection. Natural selection explains why chickadees are disposed to return to cached food locations and can do so robustly, doubtless in part through explaining why various learning mechanisms directed at getting food have been selected. So, natural selection partly accounts for the instance of these dispositions we find in this individual, Jayla, around today. This is a paradigm case: all three stabilization processes have been at work. Each separately is a basis on which the outcome of getting food is a stabilized function of the bird's behaviour. Thus, having a stabilized function does not depend on having an evolutionary history (§3.6 below). Nor need all three stabilization processes be pulling in the same direction, as they are in this paradigm case.

In sum, there are natural reasons why, in biological organisms, robust outcome functions also tend to be stabilized functions. These come together to constitute *task functions*. It is usual to talk of entities having functions, the function to produce a certain output or cause a certain outcome. The outcomes so produced are also sometimes described as functions. It will be convenient for us to adopt that (slightly strained) terminology. So, task functions are outputs produced by a system. A type of output counts as a task function if it is robust (§3.3) and has been stabilized (§3.4). Outcomes can also be robust as a result of intentional design. That forms a further, alternative basis of task functions (§3.5).

Noting that robustness and stabilization converge still leaves open the question of how an organism manages to achieve outcomes robustly. What is the synchronic mechanism by which those outcomes are produced, and produced robustly in the face of variation in conditions encountered when they are initiated and while they are being executed? What was the synchronic mechanism that keyed those behaviours **(p.51)** into conditions in which they were stabilized through survival, learning, and/or natural selection?

Task functions need not be generated by representations of, for example, conditions, goals, or targets. Developmental outcomes can be robust in virtue of a collection of parallel and backup mechanisms without any representations being involved. Nevertheless, in many cases there is an internal-components explanation of how the system achieves its task functions, an explanation that falls within our overall framework for representational content.[1] There are internal components which stand in exploitable relations to aspects of the environment that are relevant to achieving an outcome (a task function), where an internal process performed over vehicles with those properties constitutes an algorithm for achieving the distally characterized outcome successfully in a context-sensitive way.[2] That is to say, the third element in the natural cluster is

having the kind of internal organization that is characteristic of being a representational system of the kind we have been discussing. This third element of the cluster is made more precise in other chapters—in particular, Chapters 4 and 5 specify the kinds of algorithm involved.

In short, we can observe that three features tend to cluster together: producing outcomes robustly, those outcomes having been stabilized, and their being produced by a mechanism in which internal components stand in exploitable relations to relevant features of the environment (see Figure 3.1). It is the existence of this clustering that constitutes the internal components as being representations and gives representational explanation its distinctive explanatory bite. This collection of real patterns allows us to make a rich set of inferences when we recognize a system's representational properties. When we come across instances of this cluster, a whole new explanatory scheme comes into a play, a scheme which supports a host of defeasible inferences—inferences for example about ways of acquiring and weighing sources of information, of building constancy mechanisms, and of processing information optimally, to give just three examples from the host of findings catalogued by psychology, information theory and the other cognitive sciences. On one reading of the homeostatic property cluster view of natural kinds (Boyd 1991), having representational content in accordance with this sufficient condition is a natural kind.[3] Finding a system of this special **(p.52)** type tells us a lot about it, allowing us to predict and explain it in ways that would be unavailable or less perspicuous in non-representational terms.

The next two sections characterize the two aspects of task function more precisely, illustrated by a case study from psychology on the mechanisms of motor control. We start with robust outcome functions and then move on to precisify stabilized functions.

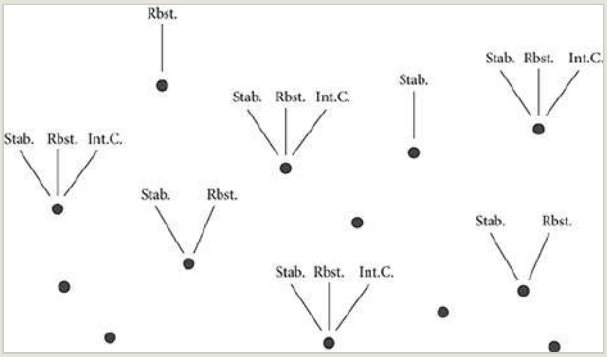

Figure 3.1 Outcomes produced by organisms/systems (black circles) can be robust (Rbst.), stabilized (Stab.) and/or produced as a result of the interaction of internal components bearing exploitable relations to the environment (Int.C.). Outcomes can have subsets of these features (labelled), or none (unlabelled black circles), however these features

tend to cluster together, and do so for a natural reason (see text).

3.3 Robust Outcome Functions

The first requirement on task functions is that they should be robust. This section develops the relevant notion of robustness. Robust outcome functions are roughly those outcomes that result from behaviour which we humans are inclined to perceive as being goal-directed. Think of the squirrel which raids nuts from a bird feeder on a pole, crawling along a thin branch, battling with the wind, losing balance and recovering, overcoming the 'squirrel-proof' collar on the feeder, and obtaining the food. It is impossible to watch the squirrel's antics without its goal seeming obvious. The tendency to see behaviour as goal-directed can in fact be activated by appropriate movements of geometric shapes, as well as human and animal behaviour. It develops early in childhood and seems to be an important precursor to an explicit understanding of others' mental states (Abell et al. 2000, Aschersleben et al. 2008, Biro and Leslie 2007, Frith and Frith 1999, Gergely and Csibra 2003). Although we tend to see them as such, not all robustly produced behavioural outcomes depend on represented goals. For our **(p.53)** purposes it is important to characterize robust outcome functions without presupposing that they are generated by goal-representations (or any other representations).

Motor control of reaching offers a paradigm example of robustly-produced outcomes. This is a useful case study for us because experimental work has delivered a detailed understanding of the mechanisms by which movements of our limbs are controlled subpersonally so as to reach their targets fluently. There is an online mechanism adjusting the action as it unfolds and a diachronic mechanism that tunes the online mechanism as a result of feedback. The online mechanism makes continual adjustments to the movement while the action is being executed. If the target is displaced, the trajectory of the limb is adjusted so that the finger still reaches the target (Goodale et al. 1986, Schindler et al. 2004, Milner and Goodale 2006). Those adjustments are made even when the target is shifted surreptitiously during a saccade, showing that conscious recognition that the target has been displaced need not be involved in this form of control (Fourneret and Jeannerod 1998, see Figure 3.2).

The diachronic mechanism tunes the online system so that it remains effective. Subjects fitted with prismatic goggles that shift all visual input 15 degrees to the left initially make mistakes when trying to touch a target, reaching nearly 15 degrees to the right. Over a series of trials their dispositions

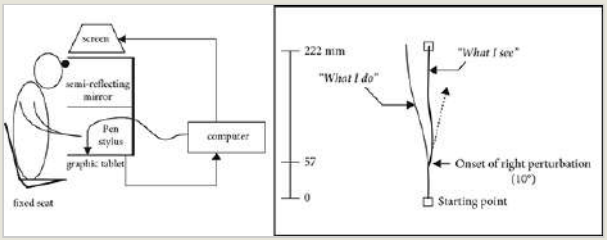

Figure 3.2 The task from Fourneret and Jeannerod (1998). Subjects adjust their reaching trajectory during action

execution even when the target is moved surreptitiously during a saccade.

adjust so that they reach the target again (Redding and Wallace 1997, Clower et al. 1996). Online guidance control remains in place, with the in-flight adjustments now being appropriate to the new set-up. When the goggles are removed, an error in the opposite direction is observed and adaptation in the reverse direction begins. Similar adjustments over time occur if there is interference at the output side by having subjects make their actions in an artificial force field (Thoroughman and Shadmehr 2000). This mechanism of adaptation recalibrates our reaching dispositions as we change and grow. Patients with damage to the cerebellum exhibit online guidance control of reaching but their behaviour does not adapt to prism goggles or an artificial force field (Smith and Shadmehr 2005, Bastian 2006).

(p.54) Motor control illustrates two key features of robust outcome functions: (i) the same distal outcome is produced in response to a variety of different inputs to the system; and (ii) the outcome is produced successfully across an array of relevant external conditions. This corresponds to the way Ernst Nagel characterized goal-directedness (his ‘system property’ view: Nagel 1977, pp. 271–6; crediting Sommerhoff 1950; see also Bedau 1992). Nagel separated out two kinds of variation across which the same outcome is produced or pursued: variations in initial conditions, and perturbations occurring during action execution. In many cases a perturbation can be considered as simply producing a new initial condition from which the organism may be able to reach the same goal. If our squirrel falls off the branch during its approach to the feeder, then its location on the ground is a new condition in which it will still be able to pursue the goal of getting the nuts. Other perturbations are external conditions that would prevent the system from reaching its goal, like the wind encountered when the squirrel is balancing along a fence. We can simply treat external circumstances encountered at the outset and during execution as specifying a complex condition under which the system may or may not reach a given target. Robust outcome functions are successful across a variety of such conditions, and the system is disposed to produce such outcomes in response to a variety of different inputs.

Some authors have proposed a further requirement for behaviour to count as goal directed: that the organism should bring about the outcome robustly by doing different things in different circumstances (Walsh 2012: selecting actions that are goal-conducive from an available repertoire). Should that requirement be built into our account of robustness? It is indeed a feature of motor control. Online guidance means that different sequences of motor output are deployed depending on the obstacles and disturbances encountered during execution (Schindler et al. 2004). The previous paragraph advanced a more minimal condition. Whether the organism produces the outcome must be input-sensitive,

and it must do so in response to different inputs. Should we also require that it should use a range of different means?

Counting against the stronger requirement, it is common for natural selection to result in a cover-all strategy, where producing the outcome is sensitive to relevant external circumstances, but the outcome is produced by just one means. For example, one way of getting a peg into a hole is to grab it with a rubbery arm that shakes indiscriminately rather than targeting the hole.[4] For a biological example, consider a plant that pursues a cover-all strategy to get a seed into a light gap in the forest: it distributes seeds indiscriminately in all directions. Natural selection will typically make that behaviour sensitive to a variety of different cues about the season, so that the behaviour is produced at an appropriate time. But the outcome is not brought about via a variety of behavioural outputs. The argument that stabilization and robustness are linked in a **(p.55)** natural cluster extends to such cases, so we should not require robust outcome functions to be produced by a repertoire of different means.

Notice that I did not say that the organism has to be targeting its behaviour on an object, still less that input sensitivity must be a matter of sensitivity to features of an object, like tracking its location. Cybernetic accounts of goal-directedness are modelled on control systems that achieve a goal by interacting with a target (our simple motor control case is like that). Cybernetic accounts do not extend easily to behaviour when there is no target object present, for example foraging for an absent foodstuff (Scheffler 1959). Our account of robust outcome functions has no such limitation. Nor is automatic behaviour excluded by this account. Automatic and stereotyped behaviour, like the frog's tongue protrusion in response to very specific fly-like visual stimuli, can in principle count, provided it is produced in response to different inputs and the behavioural outcome is achieved in a range of different external circumstances.

Nevertheless, not every kind of behaviour that is 'robust' in a pre-theoretic sense will qualify as a robust outcome function. A ball that simply shakes itself would reach the bottom of a rough shallow crater from many different initial positions. But it is not sensitive to inputs, either in when it produces the shaking behaviour nor in what kind of output it produces. The system is not in any way adapting its behaviour to its circumstances. Shaking indiscriminately in all circumstances is not the kind of behaviour that calls for representational explanation.

Taking stock, we have arrived at the following definition of when an output F counts as a robust outcome function produced by a system S. F is a type of output. S can be an individual system or a lineage of systems. In the second case 'S' picks out systems typed by the property in virtue of which they fall in the same lineage (e.g. being members of the same species). Recall that calling an

output F a function is shorthand for S having the function to produce F (in certain circumstances).

Robust Outcome Function

An output F from a system S is a *robust outcome function* of S

iff

(i) S produces F in response to a range of different inputs; and
(ii) S produces F in a range of different relevant external conditions.[5]

'Output' is a neutral term that covers bodily movements, actions, and consequences of actions.[6] As I use the terms, bodily movements can be characterized by purely intrinsic properties of the system, for example moving the eyes 12 degrees to the right is a bodily movement. Actions can be and usually are world-involving; for example, pulling a lever or moving to a particular location. Actions have consequences in the world which **(p.56)** may or may not be further actions. Getting a pool ball into a pocket is an action; winning £50 as a result is a consequence. All of these are 'outputs' caused by the agent and could qualify as robust outcome functions.

For condition (i), we need to look at the facts of a particular case to assess what counts as a different input. They have to be differences that the system is sensitive to in some way (e.g. an undetectable difference cannot count). A generalization that follows from how a mechanism deals with one type of input is not sufficient either. For example, a mechanism that triggers an internal state R when a temperature of 20°C is detected might, without further elaboration, do the same at 19.5°C and 20.5°C. Evolutionary pressures could select for this kind of stimulus generalization. Nevertheless, these values would not count as different inputs. They specify the range of values that count as an input of the same type for this kind of mechanism. If, on the other hand, R is triggered by a temperature of 20°C and also an increase in light levels, then those do count as different inputs.

The idea of different relevant external conditions in (ii) also needs careful handling. A different alignment of the planets is a different external condition but is not (usually) relevant to whether an outcome can be successfully produced. Relevant conditions are those that would impact on the system's ability to achieve an outcome or would affect whether the outcome is likely to be successful. In the seed-scattering example, the difference in where the nearest light gap is found, and hence where the seedling germinates, is a difference in a relevant external condition.

3.4 Stabilized Functions: Three Types

(a) Consequence etiology in general, and natural selection

The second element in our cluster is the category of stabilized functions. They correspond broadly to the second aspect of Aristotelian teleology: the idea that teleological outputs are produced because they lead to good consequences. In discussing the cluster (§3.2) I argued that robust outcomes tend to go along with being the target of natural selection and/or of learning, and/or with contributing to the persistence of individual organisms. This section spells out those conditions more precisely. Taken in the alternative, they define our category of stabilized functions.

How can an output be generated because of the consequences it will produce? Of course, an agent can do that. But agency presupposes intentionality. For a long time it was not clear how to account for teleological causation without presupposing intentionality. Darwin famously showed that there is no mystery. An output is generated because of the good effect it will produce when, in an organism's evolutionary history, outputs of that type have contributed to survival or reproduction. In that case the organism is producing this output now in part because of effects that the same type of output has had in the (evolutionary) past. Larry Wright generalized that idea: F is a function of S just in case (a) F is a consequence of S's being there; and (b) S is there because it does F (Wright 1973). Wright's definition covers processes **(p.57)** like feedback-based learning operating within the life of an individual organism as well as processes like evolution by natural selection operating over lineages of organisms.[7] It applies to any process where outputs in the past have had consequences that explain the current existence of a system disposed to produce outputs of the same type. I use 'consequence etiology' to cover any output which satisfies Wright's definition (see Figure 3.3).

Wright's definition has long faced the objection that it draws the category of function too broadly (Boorse 1976). For our purposes that is problematic because the definition is much broader than the kinds of stabilizing processes found in our natural cluster. It applies to a small rock that keeps its place on the river bed by holding a larger rock in place above it in a stream; also to a leaky gas hose that keeps on emitting gas because the gas poisons every person that comes near enough

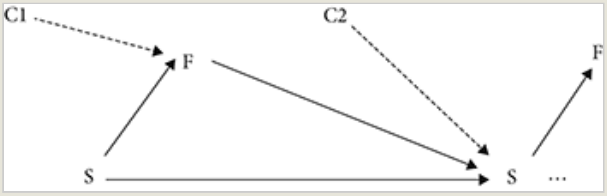

Figure 3.3 The most general treatment of consequence etiology. An output F occurs because a system S exists (right-hand side); and S exists because it, or its ancestors, have produced F in the past (left-hand side). Both causal processes may depend on certain environmental conditions obtaining (C1, C2).

to fix it. Contribution to survival of an organism is perhaps the most widely applicable kind of stabilization in our natural cluster, but even that is a special case of Wright's formula. It calls for an organism, one that acts to maintain or promote its survival in the face of changes to internal and external conditions.

Our task then is to delineate the class of stabilized functions in a narrower way than Wright, so as to coincide with the natural cluster that underpins representational explanation. Since any single cover-all condition like Wright's is liable to over-generate, I will adopt a disjunctive definition of stabilized functions. Evolution by natural selection is the first case. It is a well-understood basis for stabilized functions. I intend it to extend to cases where selection stabilizes the presence of a trait in a population but has not gone to fixation; also to selection on culturally transmitted traits. The next two subsections focus in turn on the other two kinds of consequence etiology that show up in our natural cluster: contribution to persistence of an organism; and learning with feedback.

(b) Persistence of organisms

The most ubiquitous way that natural selection has made outcomes robust is by inventing the organism: a complex system that is separated from and markedly out of **(p.58)** equilibrium with its surrounding environment, and which continually creates the conditions for its own persistence in that state.[8] By continuing to exist, organisms are able to continue to produce the types of outputs they have produced in the past, enabling robustness.

Philosophers have put forward several accounts of biological function in terms of contribution to persistence: staying alive (Wouters 1995, 2007), self-reproduction (Schlosser 1998), active self-maintenance (Edin 2008), or maintaining a differentiated organized system (Mossio et al. 2009). Christensen and Bickhard's account of functions is in this spirit (Christensen and Bickhard 2002). According to them the task towards which functions are directed is a system's capacity to generate the conditions for its own persistence when it is out of equilibrium with its surrounding environment. Functions of components of a system are Cummins-style contributions to this overall capacity.

Our stabilized functions are outputs of a whole system, rather than of components,[9] but they contribute to persistence in something like this fashion. Rather than starting with difficult concepts like self-maintenance and being out-of-equilibrium in the appropriate way, we can focus on the kind of persistence that figures in our cluster—that is, the persistence of organisms. Organisms are a special kind of self-maintaining system. They resist the tendency to disorder by maintaining a boundary, moving energy across it, and continually rebuilding themselves to keep themselves in an improbable state of differentiated organization. Godfrey-Smith uses the term 'self-production' to distinguish organisms from other self-maintaining systems like a car that monitors its states and fixes some problems (Godfrey-Smith 2016, following 'autopoiesis': Maturana

and Varela 1980). Organisms are also self-producing in a stronger sense than we find in cases like the leaky gas hose and the rock on the river bed. An account of what it takes to be an organism would open up debates about equilibrium, self-maintenance, and self-production which would distract us from our enquiry, so my definition will help itself to *organism* as a biological category. It is contribution to the persistence of an organism that should count as a stabilized function for our purposes.[10]

Chemotaxis in *E. coli* bacteria is a good illustration of the way behaviour can contribute to the persistence of an individual organism. An individual swims in a straight line, but when it detects that the concentration of one of a number of harmful chemicals is increasing, it performs a random 'tumble' to a new direction of travel (Berg and Brown 1972). The effect of this behaviour is to take the bacterium away from dangerous chemicals (often enough), thereby contributing to its persistence. Moving away from harmful chemicals is a distal outcome of the bacterium's behaviour, an outcome that contributes to its persistence. This is a typical case where robustness of the outcome (safe location) in the face of variation in external and internal biochemical **(p.59)** parameters (Alon et al. 1999) goes together with those outcomes contributing to the persistence of the individual organism.

Where an output has contributed to the persistence of an individual organism we can give a consequence etiology explanation of its current behaviour. It behaves a certain way now partly because it behaved in the same way in the past, which had consequences that kept it alive, raising the probability that there is something around now that will produce an output of the same type. We come across an instance of bacterial tumbling behaviour partly because that kind of behaviour has kept the individual bacterium alive, together with its disposition to tumbling behaviour. That is a historical rather than a forward-looking or counterfactual way of explaining behavioural outputs. Indeed, without the historical angle we would be back to the mystery of teleological causation, the mystery of how it is possible to explain a cause in terms of the type of effect it is likely to produce (without appealing to intentionality in the causal agent).

When an occurrence of F contributes to the persistence of an organism S, its effect is not specific to F. It raises the probability that any of S's outputs will be produced (since S is still around to produce them all). That is unlike natural selection, which increases the probability of organisms producing F rather than alternatives; or feedback-based learning, which specifically raises the probability of S producing F in appropriate circumstances. Furthermore, in an organism capable of feedback-based learning, contribution to persistence has the effect of keeping an organism with the disposition to F around for long enough for learning to make F-production more robust, improving its discrimination about

when to produce F or acquiring new means for bringing it about. Persistence is then an indirect route to making an outcome F robust.

(c) Learning with feedback

Returning to our motor control example, the way reaching adapts when subjects put on prism goggles illustrates the importance of learning in producing robustness. Reaching behaviour studied in non-human animals often leads to outcomes that contribute directly to the persistence of the animal performing the experiment. A macaque receives food or juice as a result of where it reaches or moves its eyes, which contributes directly to the persistence of that individual organism (Kiani and Shadlen 2009, Chestek et al. 2007). Human subjects are generally rewarded with money or course credit rather than food. In that case the outcomes produced don't directly explain the persistence of the individual organism.

Outcomes do, however, explain why a given behavioural tendency arises or persists in an individual. For example, a person in a reinforcement learning experiment might learn to press the 'F' key on a keyboard in response to some arbitrary image A and the 'J' key in response to another image B. Those behaviours are reinforced because the subject is given points that will turn into a monetary reward at the end of the experiment. If we focus on the disposition to press the 'F' key in response to image A, then an account of why that behavioural disposition exists in the individual mentions the outcomes that have been elicited by pressing the 'F' key in the recent past. Learning can **(p.60)** also explain the robustness of a behavioural disposition; for example, the ability repeatedly to touch a screen within a small target area, across small variations in the initial conditions, and in the face of noise in the perceptual and motor systems (Wolpert and Landy 2012; see Figure 3.4). There is of course also a learning explanation of the macaque's reaching behaviour. It is stabilized both by learning and by contribution to persistence.

Isn’t flexibility the converse of robustness? Learning is an interesting case because it shows how flexibility is important for robustness. We often find in biology that keeping some properties constant calls for sensitive flexibility in others. We see this in the way that motor control is constantly being retuned as the system’s input and output properties change (optical properties, weight of the limbs). Learning allows for plasticitiy in the circumstances in which, and means by which, an outcome in produced—leading stabilized outcomes to be produced with greater robustness.

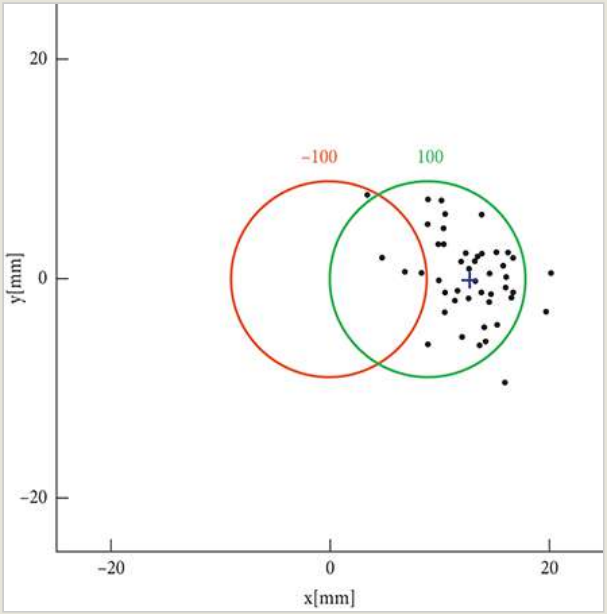

Figure 3.4 The rapid reaching task from Wolpert and Landy (2012). Subjects gained 100 points for touching the screen within the right-hand circle (shown to subjects in green) and lost 100 points for touching in the left-hand circle (shown to subjects in red). Touching in the overlap therefore produced zero points. People mainly touch within the most rewarding area because they learn from feedback how to reach (observing previous outcomes).

Learnt behaviours have evolutionary functions that derive from the function of the learning mechanism (Millikan 1984). Humans readily learn to recognize conspecifics by their faces. Human infants look preferentially at faces, which allows them to learn the statistical patterns that are indicative of face identity (Johnson et al. 1991). If we **(p.61)** suppose for a moment that no social feedback is involved, then the reason the infant acquires a new behavioural disposition—for example, to track a new person A as they come and go—does not depend on any feedback the individual infant has received. The function of the mechanism is indeed to track person A, but that is an evolutionary function, deriving from the (plausible) evolutionary function of the learning mechanism, namely to track conspecifics by their faces. This is a case where evolutionary functions do deliver quite specific stabilized functions for the products of learning.

In other cases derived functions are much less specific. Classical conditioning is a very general learning principle. It enables organisms to reidentify statistical patterns in the inputs they receive. When an association has been learned, what is it supposed to track? The evolutionary function of the learning mechanism only tells us something very general. Its function is to track something useful

that correlates with patterns in the input. Once a new association is put to use to condition behaviour, if that behaviour is stabilized then feedback-based learning may underpin a much more specific function, as we will see in a moment. But before being connected up to behaviour, the functions of a new association derive only from the evolutionary function of classical conditioning and are highly indeterminate. Basic sensitization, where a response is attenuated when a stimulus is repeated, is another case where the mechanism of behavioural plasticity has only a very general-purpose evolutionary function.

When an organism's behavioural dispositions are modulated by feedback, learning underpins stabilization directly, irrespective of any evolutionary function. Feedback need not be in the form of a commodity that has an evolutionary function (a primary reinforcer). People will shape their behaviour to feedback in the form of money, or the promise of money, or tokens that will be exchanged for money; also for positive social feedback; and so on. A stabilization-based explanation need not descend to an explanation of why monetary feedback stabilizes behavioural dispositions. If an agent's behavioural dispositions are in fact stabilized by a variety of outcomes O_i, then we can explain a current behavioural disposition (e.g. to touch the region inside the green circle on the computer screen) by the fact that outputs of this sort in the recent past produced one such outcome, O_1 say. It is then a further question as to why O_1 reinforces behavioural dispositions in that agent.[11] An answer to that question need not form part of a stabilization-based explanation of why the agent has a given behavioural disposition now (e.g. to touch inside the green circle).

Like natural selection, reinforcement can lead an organism to produce O more robustly by better detecting the circumstances in which its behaviour is likely to produce O; by adopting new means for producing O in new circumstances; or by increasing the robustness with which it can produce a particular means to producing O. Learning is more sophisticated than natural selection in some respects. One-shot learning is **(p.62)** possible in some cases. Then a single episode in the past explains why an individual has a behavioural disposition now. Nearby outcomes can be reinforcing. Where O is the target of learning, achieving an outcome that is close to O (along some relevant dimension) can make it more likely that the organism will achieve O on the next occasion. That is, outcomes that are closely related to O can contribute to likelihood that O will occur in the future. Where an outcome comes in degrees, like the quantity of juice received, the organism may shape its behaviour so as to increase the quantity delivered. Negative reinforcement is also common, for example a rat forced to swim in a Morris water maze will learn how to behave so that it has to swim for less time in the future. So, the stabilized function is getting to a submerged platform (O_1) and the feedback which explains its stabilization is the unpleasant effect of not reaching the platform ($\neg O_1$). In these two cases, it is not

producing O itself but producing outcomes closely related to O that has contributed systematically to the organism’s disposition to achieve O.

Learning by imitation is an interesting case. It takes several forms. Sometimes it is driven by social feedback; for instance, that people smile or give other signs of approval. That is a case of reinforcement and fits into the characterization we have just given. Or the learning may occur because the individual performs the behaviour and receives some other kind of reinforcing feedback, like food or warmth. In other cases people may acquire a behavioural disposition without feedback, just because they see others perform it. That disposition will not then have been stabilized through feedback-based learning, but there may well be another stabilization-based explanation; for example, the behaviour which is transmitted may have been stabilized in the person’s lineage or social group through cultural evolution.

It would take us too far astray to catalogue all the types of learning and to give an account of what characterizes the different kinds. For our purposes it is enough to point to the category of feedback-based learning, as used in the behavioural sciences, and to note that it is a strong form of stabilization that tends to go with robustness in our natural cluster.

(d) A ‘very modern history’ theory of functions

This section constructs my notion of stabilized function out of natural selection, learning, and contribution to persistence, and defends its historical character.

It would be handy if we could treat stabilization synchronically, on the model of forces that are holding a system in place. But dispositions an organism could exercise are not like forces or other outputs that are operating continuously. Our stabilized functions are not like the kinematic equilibria studied in physics. That makes it tempting to adopt a counterfactual or forward-looking approach. Stabilized functions would then be outcomes that would be stabilized were they to be produced, or that are likely to be stabilized in the future.

The difficulty is that it is a very open-ended matter whether an output would contribute to the persistence of an organism, or would be stabilized by feedback-based learning, or would promote reproductive fitness. All outcomes that would contribute to **(p.63)** the persistence of an individual would count as being amongst its stabilized functions. However, whether an outcome will contribute to persistence is a notoriously open-ended matter. It depends heavily on the context. And within a context, whether a behaviour will in fact be stabilized will depend upon accidental features of the process that ensues. Outputs that seem very unlikely to contribute to persistence might end up doing so through a series of compensatory accidents (as happens to the cartoon character Mr Magoo). Without some other constraints, there are just too many effects that would be stabilized in some circumstance or other, hence too many functions. Facts about

what could contribute to persistence are much more open-ended than historical facts about what has actually contributed to the persistence of an individual organism. The same is true for natural selection and for learning.

A second strong reason not to rely on a forward-looking form of stabilization is that such functions are of the wrong kind to figure in causal explanations. Recall the mystery of teleological causation, namely of understanding how a good effect could 'draw out' a cause suited to producing it. Wright's way of making that un-mysterious, and Darwin's, is to point to consequence etiology, in which functions are a matter of the effects that such outcomes have produced in the past. If we seek to explain why a system produced an outcome O, it is unilluminating to cite the fact that O is likely to be stabilized in the future (i.e. to cite a future-directed function). Functions based on stabilization history can figure in an explanation of why outcomes are produced; forward-looking functions cannot (not straightforwardly). A historically based approach to function thus has better credentials to figure in casual explanations than a forward-looking approach to function does. Any explanatory purchase of forward-looking functions would proceed through a historical generalization—that they tend to have been the result of some stabilization process. Furthermore, our functions need to connect with the natural cluster that underpins representational explanation. It is actual historical processes of stabilization that appear in that cluster.

What is relevant for our purposes, then, is contribution to stabilization through: natural selection over a lineage of systems, learning within an individual system, or persistence of an individual organism. Teleosemantics standardly appeals to the first two (although with some problems with the way learning is incorporated). I expand the category to accept the widespread suggestion (e.g. Christensen and Bickhard 2002) that functions can be a matter of contribution to the persistence of self-producing systems (for our purposes, organisms). I follow Godfrey-Smith's insight that appealing to actual causal history is the right way to cut down on the problematic liberality of forward-looking accounts of function (Godfrey-Smith 1994b). Godfrey-Smith calls his appeal to the most recent evolutionary function of a trait a 'modern history' theory of function. We could then call ours, which includes the recent learning and persistence history of an individual organism, a 'very modern history' theory of function. These functions can arise just from the history of an individual organism, including very recent learning and contributions to its persistence, irrespective of any history of selection.

(p.64)

Stabilized Function

An output F from a system S is a *stabilized function* of S

iff

producing F has been systematically stabilized:

(i) by contributing directly to the evolutionary success of systems S producing F; or
(ii) by contributing through learning[12] to S's diposition to produce F; or
(iii) where S is an organism, by contributing directly to the persistence of S.

The evolutionary condition is deliberately drawn so as to cover cases of cultural transmission, which may have been important in human cognitive evolution and thus in generating representational content in many aspects of human psychological systems (Sterelny 2015). It also covers cases where selection has been operative but has not gone to fixation.

A system's behaviour will generally result in a causal chain of outputs, which can vary in robustness along the chain. Stabilization homes in on only one or a small number of steps in this chain. When a macaque moves its arm to pick up a grape, getting the grape and moving the arm both make a causal contribution to that form of behaviour being stabilized, and to the individual persisting, but only getting the grape does so directly. On the other hand, idiosyncratic things may have happened in the individual's history. Where a behavioural episode accidentally happens to produce some beneficial or reinforcing effect, and there is no systematic explanation of why that is so, then it does not even start to generate stabilized functions, even if the episode made some contribution to the persistence of the individual or the chance that it would produce a particular kind of behaviour in the future.

3.5 Task Functions

This section puts the pieces together and defines task function, which I argue is the right account of function to figure in accounts of content in our case studies. Task functions combine stabilization with robustness. There is a source of robustness which we have not yet considered, namely deliberate design. A human can design a system to perform a task; that is, to produce certain outcomes robustly in certain circumstances. Design need not involve any history of stabilization. Indeed, artefacts can be designed to produce an outcome robustly which would not be stabilized by feedback. For example, we could design a robot that would navigate to a power source and use the energy to blow itself up, with the ability to do so robustly from a variety of starting points via a variety of routes. So, we need to include design functions as an alternative to stabilized functions.

(p.65) Task functions based on design do not meet our criteria for naturalism. What a system has been intentionally designed to do depends on the mental states of the designer, so this is not a non-semantic, non-mental source of functions. It does not form part of our account of where underived content

comes from. Nevertheless, it is worth recognizing the products of design as having task functions since they pattern with the other cases, before setting aside design in order to focus on the underived cases.

The kind of design we include is where a person has designed a system to produce certain behavioural outputs. Another way content can arise derivatively is more direct: a person can intend a representation to have a certain content. A sentence may mean what its writer intends it to mean. A computer database may represent what its programmer intends it to represent. Directly derivative content does not depend on task functions at all. It comes directly from the user's intentions or beliefs about what a vehicle represents, so our definition of task functions does not cover these cases.

So, an output is a task function if it is robustly produced and is the result of one of the three stabilizing processes discussed above or intentional design:[13]

Task Function

An output F from a system S is a *task function* of S

iff

(a) F is a robust outcome function of S;
and
(b)
(i) F is a stabilized function of S; or
(ii) S has been intentionally designed to produce F.

I am not suggesting this as an analysis of biological function. Some have argued that representing correctly is genuinely normative, and that biological function is genuinely normative, and that the otherwise puzzling normativity of content can be dissolved by showing that it reduces to the normativity of biological function. It will be apparent that I am not engaged in that project. Both biological function and (subpersonal) representational content are descriptive categories (see §6.5). My definition of task function does however have several features that are familiar from biological functions. A system can have task functions that it is no longer able to perform. It can malfunction, which is different from not having a function at all. And it can produce outputs which are side effects, which regularly accompany task functions but have not been the target of stabilization or are not robustly produced.

Nor do I claim that task function is the only notion of function to which representational content can be connected (recall the pluralism). My claim is that task functions are suited to giving an account of representational content in many kinds of subpersonal psychological system. Task function is a necessary part of some sufficient conditions for content. (Since the definition of task

function is disjunctive, and the **(p.66)** definition of stabilized function is disjunctive, it really generates several different sufficient conditions.)

Task functions are part of a natural cluster, a real pattern in nature that I will argue gives representational content better ways of explaining behaviour than would be available otherwise (§3.6, §8.2). However, task functions can vary in ways that affect the explanatory bite of the contents they generate. This is what we should expect in biology. Robustness comes in degrees. The more robust, the more explanatory value a task function is likely to have. Similarly, stabilization comes in degrees, from powerful or long-standing stabilized functions to marginal cases; for example, evolution where there has been some selection, but no fixation or other endpoint has been reached.

Another dimension of variation lies in the various bases of stabilized functions (clauses (i)–(iii) of the definition). In a paradigmatic case an outcome like obtaining food has been the target of natural selection, and learning from feedback, and has contributed to persistence of the individual organism. But these may not line up. A rat that learns how to press a lever to deliver electrical stimulation directly to reward centres in the brain acquires new task functions (by (ii)), but ones that are evolutionarily disadvantageous (clause (i)) and do not contribute to the animal's persistence (clause (iii)). In natural cases there will normally be connections between the three stories; for example, money becomes a positive reinforcer partly because of its connection with reinforcers like social feedback for which we can give a more direct evolutionary explanation. When they dissociate, there will still be task functions, but there may be different task functions underpinned by the different clauses, and they may pull in divergent directions (see §3.7). Representational explanation will have greatest merit in the paradigmatic cases and less merit in these more marginal cases. Less paradigmatic task functions can underpin genuine representational content—they are not merely cases of 'as if' content—but if the penumbral cases were the only ones that existed in nature it would be unlikely that representational content, of the kind we have identified here, would be an important explanatory category. The marginal cases are not what makes our natural cluster explanatorily powerful, but they do get carried along for the ride.

When a property applies more widely, that is generally of explanatory benefit, but it trades off against the fact that more generally applicable properties tend to support fewer inductions. Although we can classify very many entities as *physical object under 10 kg*, falling into the category tells us little about other properties an object is likely to have—it supports few inductions. The merit of our cluster is that, as well as being found widely in nature, it supports a rich set of inductions. Robustness also gives us generality by grouping a range of different local properties together (§3.6, §8.2). A system's reaction to light and sound might be lumped together because they are both means for tracking a distal property like distance. What can look like a variety of different processes,

if we considered only the local operation of the system, exhibit commonalities when treated in terms of task functions. This generality is not achieved at the cost of reduced inductive potential (as with *object under 10 kg*), since task functions key into our cluster and a rich set of world-involving inductions about the system's interaction with distal features of its environment.

(p.67) 3.6 How Task Functions Get Explanatory Purchase

(a) Illustrated with a toy system

In this section we look at a stylized toy system that captures some essential features of the mechanisms of motor control. It will illustrate why task functions underpin a proprietary explanatory role for content. Well-confirmed accounts of motor control appeal to a variety of interacting internal components including forward models, inverse models, and comparator circuits (Desmurget and Grafton 2000, Battaglia-Mayer et al. 2014). The most basic kind of comparator circuit compares visual or proprioceptive feedback about the location of a limb with a specification of a target location in the same code, using the discrepancy between the two to update the motor program driving the limb (Wolpert and Ghahramani 2000). In this way the limb's position is adjusted until the difference between its location and a target location is reduced to zero.

Smooth control of action also depends on making internal predictions of the likely effects of executing a motor command, and adjusting the motor command in response to discrepancies between the prediction and the target state, even before feedback from the world has been received (Wolpert et al. 2011, Bastian 2006). Since I can illustrate how representational content gets explanatory purchase even without these additional internal components, I will work with a simple toy model that only contains the first comparator circuit, the one based on external feedback. Figure 3.5 illustrates this toy system S. It moves in just one dimension, along a line. From a range of initial conditions, it will move along the line until it reaches location *T*, where it **(p.68)** stops. If it is blocked or displaced along the line it will continue to move towards *T* when released. Reaching *T* is a robust outcome function of the system.

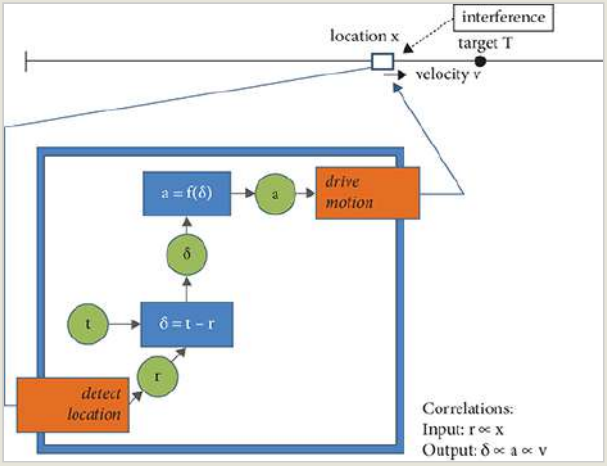

Figure 3.5 A system deploying a simple comparator mechanism.

We can explain how the system achieves the outcome of reaching *T* by appeal to its internal organization and the relations that those internal components bear to features of the environment. S has an internal register **r** that correlates with its distance from the origin, and another internal register **a** that correlates with the velocity of its wheels. A third internal state **δ** correlates with the distance of the system from *T*. That correlation is achieved by subtracting the activity of **r** from another fixed level of activity **t**. A monotonic transformation from this difference signal **δ** to **a** is such that the motion produced in the wheels drives S from any starting position to *T*, where it stops.

Reaching *T* is a distal outcome produced by S robustly using a variety of motor outputs—ways of changing the velocity of the wheels over time. What these different patterns of motor output share is that they all achieve the distal outcome of reaching *T*. Similarly, at input, S will reach *T* from a variety of different starting positions, and in the face of a number of ways of displacing S while it is executing its action sequence. So, reaching *T* satisfies the definition of being a robust outcome function of S. (The robustness is not very great, and so representational explanation will not deliver very much additional explanatory purchase, but the case is sufficient to illustrate the point.)

To get stabilized functions into the picture, we have to supplement the case. Suppose the system needs to recharge its batteries periodically if it is not going to stop moving altogether. There is a power source at *T*. Now, if we encounter S, moving around, with its disposition robustly to reach location *T*, there is a consequence etiology explanation of why. Reaching *T* in the recent historical past has contributed to the persistence of the system, with its disposition to reach that very location. We could also add a learning-based stabilized function. Suppose the internal state **t** is reset periodically at random, leading the system robustly to move to a new location; also suppose that when the system manages to recharge (by chance at this stage), that fixes the state of **t**. In future it will then robustly move to the recharge location *T*. Getting to *T* has then become a learning-based stabilized function. (We could add in the further capacity to adjust its dispositions over time in response to perturbations in the input and output mechanisms, as in the case of motor control—prism goggles and artificial

force fields—a stronger form of learning which would produce greater behavioural robustness.) Either way, reaching *T* becomes a task function.

We now have all the elements in place to explain the system's behaviour using the standard explanatory grammar of representational explanation. S has various internal components that correlate with distal features of the environment (correlation being the relevant exploitable relation in this case). **r** correlates with S's distance from the origin and **δ** correlates with S's distance from *T*; **t** with the location of a power source. There are internal processes that transform **r** into **δ**, and **δ** into **a** and the velocity of its wheels. Given the way **r** and **δ** correlate with external features of the environment **(p.69)** (S's distance from the origin and from *T*, respectively), these internal transformations constitute an algorithm for performing the distally characterized task of reaching *T*.

Now consider a particular episode of behaviour: the system is displaced and moves back to *T* where it recharges. How so? Because **r** and **δ** correlated on that occasion with S's distance from the origin and from *T*, respectively. The next chapter shows how correlations are content-constituting when they explain task functions in the right way. So, the story I have just told effectively shows how the behaviour of successfully reaching *T*, the location of a power source, is explained by **r** and **δ** representing correctly. Conversely, suppose noise affects the input system and S stops at some other location *T*′. Then we can explain failure of the system to reach *T* in terms of misrepresentation by **r**. Similarly, misrepresentation by **δ** or **a** would explain unsuccessful behaviour. This pattern exemplifies the characteristic explanatory grammar of representational explanation: correct representation explains success and misrepresentation explains failure.

(b) Swamp systems

To see why this really is a case of success and failure, consider a 'swamp' system, like S, but one that has assembled itself by chance when an earthquake struck an engineer's workshop. This swamp system would be disposed to move up and down a line along the work bench, stopping when it reached some location *T*. Reaching *T* would be a robust outcome function of the system. But now consider what would happen if a random event affected internal register **t** so that S became disposed to reach a different location *T*′ (and to do so robustly). Would that count as a failure, to be explained by misrepresentation? Or would it count as a success—success in achieving the system's new function of reaching *T*′, to be explained by correct representations (with different contents)? There is nothing yet in the picture that allows us to answer that question one way or another.

If we now add that there is a power source at *T* and we observe the swamp system a short time after the earthquake, when it has had a chance to move around and recharge, then we do have something in the story which underpins a notion of success and failure for that system. It is part of an explanation as to why that particular swamp system is around, with its disposition to reach *T* robustly, that it has reached *T* in the recent past, which has contributed to the persistence of S and its behavioural capacities. If noise now messes things up so that **δ** is calculated differently and it no longer reaches *T*, but robustly reaches *T'*, that is a failure, to be accounted for by misrepresentation at **δ**.

The foregoing is effectively an argument from intuition. It trades on the intuition that there is no substantial distinction to be made between success and failure in a system that has robust outcomes but no history (hence no stabilized functions from persistence, learning, or evolution). That won't do for our purposes. We could simply define robustly produced outcomes as successes and other outcomes as failures. **(p.70)** However, the cluster we have identified means that there is a deeper logic behind these intuitions. Robust outcomes, when not accidental or due to external constraints, are often explicable in two ways at once: both why and how they are produced. *Why* is explained by history, involving a consequence etiology, here contribution to stabilization. *How* is explained synchronically, by internal components and exploitable relations. (This is fleshed out properly in Chapters 4 and 5.) Past episodes of stabilization can explain both how robust outcomes are produced now, and why the system has a mechanism suited to producing those outcomes robustly. It is the combination of these elements which makes it the case that certain outcomes are successes and others are failures. Our intuition about the absence of success and failure in the swamp system before it has interacted with the world reflects the fact that the cluster of elements that give representational content its explanatory purchase is absent in that case.

How does this argument transfer from our toy system to organisms? Consider the motor system of a swamp macaque, produced at random by lightning striking a swamp. From the moment of creation it would have the same robust outcome functions as a regular monkey. So, if it sees a grape, the swamp monkey will grasp it and eat it. Consider also a second swamp macaque who just happens to have a robust disposition to grasp 15 degrees to the right of any grape it sees. At the moment of creation there is nothing present that underpins a substantial sense in which one swamp monkey is getting it right and the other getting it wrong. However, as soon as they have had time to interact with the world for a while, there is an important difference. In one the disposition to reach and grasp grapes has contributed causally to its disposition to behave in that way—it has been the target of learning and has contributed to persistence—in the other the disposition to reach 15 degrees to the right of grapes has not. At the moment of creation neither monkey exemplifies the cluster of properties that underpins the explanatory purchase of representational content.

Correspondingly, there is no substantial sense in which either one of them is getting it right or getting it wrong. Once they interact with the world, one monkey starts to exemplify the cluster, with which comes a substantial distinction between correct and incorrect behaviour; the other does not.

These thought experiments with swamp systems are not offered as intuitive evidence in favour of my account of task functions, but rather to illustrate the consequences of the theory. Systems which are not the result of deliberate design and have no evolutionary history, but do learn systematically from feedback, would begin to acquire task functions after a short period of interaction with their environment. The same goes for organisms whose actions contribute to their own survival. This illustrates the fact that functions based on current properties of the system (robust outcome functions) and recent causal contributions to learning or persistence (stabilized functions) can underpin representational content in a way that is independent of facts about why the system was designed or its deep evolutionary history. That would be so even for a system that does in fact have an evolutionary history—the swamp system thought experiment just serves to dramatize the fact that even in natural cases, stabilized functions can arise in a way that does not depend on evolutionary history.

(p.71) We can see that in the case of learning. Think of a child that learns to clap, based on social feedback from a parent. It produces outputs (ways of clapping) that make the parent smile and learns to perform the behaviour in appropriate circumstances (e.g. not at dinnertime). These outputs now have a stabilized function F: to make the parent smile. The child’s behaviour has that function irrespective of any facts about evolutionary history. Non-evolution-based stabilized functions are acquired gradually as an organism interacts with its environment and receives feedback that reinforces behaviour or contributes to survival. A swamp system would have no task functions at the moment of creation, but would acquire them piecemeal, and would soon have task functions, functions keying into aspects of the environment it has interacted with. As soon as a swamp system has some interactions with the environment, then there will be an explanandum at which content-explanation can be addressed (success and failure), and the system will begin to have contentful states.

Task functions are still partly historical, so I have to bite the bullet and accept that a swamp system has no contents at the moment of creation. As I have argued, however, that is the right result. In these subpersonal systems, at the moment a swamp system is created there would be no explanandum for content-based explanation to address. However, the bullet is a lot more digestible than that confronted by standard teleosemantics, which accepts that a system with no

evolutionary history would have no contents even after a long life of interacting with and learning about its environment.

We can also see how the robust outcome aspect of task functions contributes to the proprietary explanatory purchase of representational content (see also §8.2b). Because reaching *T* meets the conditions on being a robust outcome function of S, there are world-involving distal patterns involving S that are less perspicuous when we consider S's behaviour only in terms of proximal sensory stimulation and proximal motor output. The same location is reached across a variety of patterns of perceptual input. Despite the simplicity of this toy system, there are real patterns in the way S interacts with distal features of its environment that generalize across proximal inputs. (In paradigm cases there will also be generalizations across proximal outputs, with multiple motor outputs eventuating in a common distal outcome, as discussed in §3.3 above.) Explanations of S's behaviour would look more complex and disjunctive if we did not recognize those patterns.

Contrast the case of the rifle firing pin (§2.2). The pin does not enter into any patterns involving distal features of the environment that are not perfectly matched by the proximal causal story. Movement of the trigger corresponds to movement of the pin corresponds to ignition of the primer, explosion of the propellant, and discharge of the bullet. Robust outcome functions 'bridge' to common outcomes across a range of different proximal conditions. That is absent in the case of the firing pin. (This is spelt out more carefully in §8.2b.)

Notice that standard teleosemantic accounts of content require a consequence etiology but do not require robust outcome functions. That misses out on an important element of the cluster that gives representations their explanatory bite. The honeybee **(p.72)** nectar dance has evolutionary functions irrespective of any feedback that comes from collecting nectar. That qualifies as an (evolutionarily based) task function, but only if the distal outcomes (arriving at distant flowers and collecting nectar) are also robust outcomes. As a matter of fact, bees do rely on a variety of inputs before performing the waggle dance, and they do reach foraging locations robustly in the face of obstacles and variations in wind speed (Srinivasan et al. 1996). As far as I know there are also robust outcome functions in the other central examples relied on by Millikan. So, the cases do fall into a cluster that supports representational explanation. But Millikan's definition of function does not include a condition that functions should be outcomes that are robustly produced. To characterize the functions that underpin representational content in the honeybee nectar dance, and other cases of directly evolved animal signalling, we need to combine evolutionary functions with robust outcome functions for the same outcomes.[14]

3.7 Rival Accounts

Griffiths (2009) argues that analysing functions in terms of contribution to persistence delivers the wrong result in many cases (see also Artiga and Martinez 2016). Organisms have many phenotypes that are detrimental to their own survival and only make sense in terms of their contribution to fitness. Behaviour that promotes mating or feeds offspring at the expense of the individual's wellbeing are obvious examples. Griffiths's example is the heavy investment in a single mating season made by males of several species of small Australian marsupials, which greatly increases their risk of death (Bradley et al. 1980, Diamond 1982). An extreme example is the way some male spiders engage in mating despite the fact that they will end up being eaten by their female mate (Andrade 1996, Forster 1992).

No doubt there are lots of these cases in nature, with many involving representation: signalling between organisms (e.g. to achieve mating) or internal representations (e.g. of the conditions indicating that now is the time to pursue a mate). Contribution to persistence cannot help to underpin representational content in these cases. Our pluralist framework will cover such cases if the behaviour produced has evolved directly by natural selection. As with animal signalling discussed earlier, representation in these cases will be underpinned by a task function which conjoins robust outcome function with evolution-based stabilized function.

Griffiths makes a rival proposal. He has a forward-looking evolutionary approach. Functions are causal role functions that will contribute to reproduction of the organism (Griffiths 2009, p. 25). This is similar to Bigelow and Pargetter's earlier proposal **(p.73)** that functions are effects that give an organism a propensity to succeed under natural selection (Bigelow and Pargetter 1987).[15]

Unfortunately, the two objections made earlier to forward-looking accounts of contribution to persistence (§3.4d) are also decisive objections to forward-looking accounts of contribution to fitness. Whether an effect will contribute to fitness is heavily dependent on the context (of other organisms and the rest of the environment). Either evolutionary history comes back in to specify the relevant context (the one organisms of that type have evolved to deal with) or there are just too many effects that would contribute to fitness in some circumstance or other. Without relying on history, there is also considerable open-endedness about what should count as the system. This open-endedness is a good reason why accounts of evolutionary functions should be based on actual evolutionary history, not possible future or counterfactual contributions to fitness (Godfrey-Smith 1994b, Artiga 2014b). Nor is there any in-principle answer to the question of how we should count fitness prospectively (at the first generation, the second generation, or further).[16] A forward-looking approach also makes functions unsuited to figuring in a causal explanation of why an

organism behaves as it does, as argued earlier in relation to forward-looking contributions to persistence. These considerations make forward-looking evolutionary functions inappropriate as a basis for representational content.

Griffiths's examples of behaviour that promotes fitness but is bad for persistence of the individual can be understood in terms of the (historically based) evolutionary function of the behaviour. This does mean that there will be cases where the two different approaches to function are pulling in different directions in the same organism. Representations involved in the spider's mate-approaching behaviour get their content in virtue of achieving a task function based on the way behaviour of that type has promoted offspring-production (hence fitness) in its ancestors. At the same time representations involved in the spider's homeostatic mechanisms can get their content from contribution to persistence, and also in virtue of having been reinforced by some basic learning mechanisms, both irrespective of their evolutionary functions (although in this case they are likely to have evolutionary functions as well). Intentional design can also produce task functions that conflict with evolutionary-based task functions. For example, by design we could use a glow worm as a light-sensitive switch to turn on the heating when it gets dark. So, our framework allows for task functions based on evolution which do not contribute to persistence (Griffiths's case), and also **(p.74)** for task functions based on learning or contribution to persistence that have conferred no reproductive advantage.

3.8 Conclusion

In this chapter we have been examining one of the two key elements of the framework introduced in Chapter 2: the task being performed by a system. What counts as a system's tasks or functions—the functions whose performance is to be explained representationally? Answering that question is usefully constrained by the desideratum that an account of content should show why adverting to representational content allows better explanations of behaviour than would be available otherwise. Representation in many subpersonal systems forms part of a real cluster in nature, in which three elements are instantiated together, better than chance and for a natural reason. This cluster is what gives representational content its explanatory purchase. A central element in the cluster is a system's having a stabilized function: producing outcomes that have been stabilized by evolution, learning or contributing to the persistence of the organism that produces them. Stabilized functions tend also to be robust outcome functions, and the converse. The third element is that there is an internal mechanism which accounts for these outputs being stabilized and produced robustly, a mechanism in which internal components (representations) stand in exploitable relations to relevant features of the distal environment. In these cases we can see both how and why robust outcomes are successfully produced. The internal components are how, and the stabilization process is why. When the three elements are instantiated together, a sufficient condition for having

representational content is met, and recognizing such contents affords better explanations of the behaviour of the system than would otherwise be available.

Notes:

([1]) The idea that the functions or capacities of a system can be explained through causal decomposition is familiar from Cummins (1984). Unlike our task functions, which are outputs of the system of interest, Cummins functions are activities of components, each playing its role in one of these causal decompositions. Any capacity of a system is a candidate for analysis, so Cummins functions are very liberal. Without a principled way to identify privileged capacities of the system, the resulting theory of content is correspondingly liberal (Cummins 1989, 1996), contra our desideratum.

([2]) Neander (2017) advances a theory of content based on contributions of components in a functional decomposition. Unlike Cummins, Neander does identify privileged capacities that call for such an explanation (e.g. the toad's prey-capture capacity). Contents are fixed directly by teleofunctions of components, e.g. a function to respond to small dark moving objects of a delimited kind in the environment, see §6.2h.

([3]) With Boyd, I reject the need for an underlying essence that explains why these features go together. (The explanation is the one we have seen.) However, I don't take this core set of features to be flexible. My account requires all three features to be present. The many other properties that often go along with being a representer are however more open-ended and flexible, as with other homeostatic property cluster views of kinds. See also §8.2.

([4]) Thanks to Andy Clark for the example.

([5]) 'S produces F' must be true with some nomological modal force. I remain neutral on whether this should be cashed out in terms of dispositions, capacities or in some other way.

([6]) It could in principle cover any kind of effect, e.g. releasing a hormone, although movement is involved in all the cases we will consider.

([7]) E.g. it covers all the various kinds of dynamics studied in Skyrms-Lewis signalling games: replicator dynamics (with and without mutation), simple reinforcement learning, Roth-Erev reinforcement, Bush-Mosteller reinforcement, etc. (Skyrms 2010).

([8]) Reproduction of entities that don't count as organisms/autopoetic systems is in principle possible, although there is debate about whether there was actually such a stage in the origin of life (Martin 2005).

([9]) See §3.3 (unless those components count as systems in their own right).

([10]) In what follows ‘persistence’ is always persistence *of an organism*, even when I omit the qualification for the sake of brevity.

([11]) That need not be because the organism can represent the reinforcer. Nor does the learning-based explanation of an organism’s behavioural dispositions presuppose that learning depended on representations (of reinforcers or outcomes).

([12]) As discussed in §3.4c, this is intended also to cover nearby reinforcement, where producing an outcome close to F (along a relevant dimension) accounts for S’s disposition to produce F; also negative reinforcement, where the disposition to do F has been stabilized by the negative consequences that have flowed from doing things contrary to F.

([13]) Usually the contexts in which the output is produced robustly will largely coincide with the contexts in which it was stabilized.

([14]) Shea (2007b) argued that a similar move addresses the dormitive virtue problem with teleosemantic contents.

([15]) Nanay (2014) makes a related proposal: that the functions which teleosemantics should rely on can be analysed in terms of subjunctive conditionals about fitness: effects that would contribute to fitness of the organism.

([16]) Standardly, fitness is measured in terms of expected long-run genetic contribution to the population, but whether that is the best measure for predicting evolutionary change over time will depend on the particular situation.