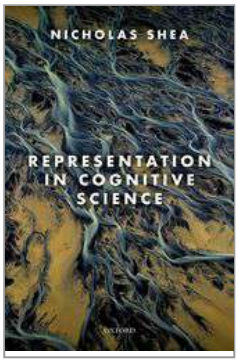


Representation in Cognitive Science
Nicholas Shea

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

Correlational Information

Nicholas Shea

DOI:10.1093/oso/9780198812883.003.0004

Abstract and Keywords

Correlation is the first exploitable relation we will consider. Correlations turn into content when they are exploited by a system: the content-constituting correlations are those which unmediatedly explain a system’s performance of its task functions (and thereby qualify as UE correlational information). This chapter shows that this approach works for fixing content in a range of case studies from cognitive science. It does so without having to appeal to representation consumers whose outputs play a content-constituting role. In each case study, contents fixed in this way do a good job of underpinning the characteristic explanatory grammar of representational explanation: correct representation explains successful behaviour and misrepresentation explains failure.

Keywords: exploitable relation, correlation, information, UE information, consumer, hidden layer, corollary discharge, analogue magnitude, plaid motion, probabilistic representation

4.1 Introduction 75
(a) Exploitable correlational information 75
(b) Toy example 80

4.2 Unmediated Explanatory Information 83
(a) Explaining task functions 83
(b) Reliance on explanation 88
(c) Evidential test 89

4.3 Feedforward Hierarchical Processing 91
4.4 Taxonomy of Cases 94

4.5 One Vehicle for Two Purposes 96
4.6 Representations Processed Differently in Different Contexts 97
(a) Analogue magnitude representations 97
(b) PFC representations of choice influenced by colour and motion 100
4.7 One Representation Processed via Two Routes 103
4.8 Feedback and Cycles 106
4.9 Conclusion 110

4.1 Introduction

(a) Exploitable correlational information

Chapter 2 introduced a framework for understanding representational content. Chapter 3 filled in one half of the framework: the functions being performed by an organism or other system. The other half is having an internal organization that capitalizes on exploitable relations—relations between internal states and the world that are useful to the system. Not all task functions are achieved representationally. Representation arises where a system implements an algorithm for performing task functions. That in turn has two aspects: internal vehicles stand in exploitable relations to features of the environment which are relevant to performing the task; and processing occurs over those vehicles internally in a way that is appropriate given **(p.76)** those relational properties. Content is constituted in part by exploitable relations: internal processing implements transitions between vehicles which make sense in the light of these relational properties, transitions called for by an algorithm which is suited to producing the input–output mapping given by the system's task functions. This chapter focuses on cases where correlation is the candidate exploitable relation. The next chapter looks at structural correspondence.[1]

The account shares with teleosemantics a reliance on teleofunctions (Chapter 3) and the insight that the way a representation is used downstream is important to fixing its content (for me, also the way it is produced). However, we will see that my account does not presuppose that there are dedicated representation consumers that play a special role in constituting content. That is an advantage of my view over some standard teleosemantic treatments (§1.4, §1.5).

An object or process carries correlational information just in case one or more of its properties correlates with the properties of some other object or process. More formally:

Correlational Information

Item(s)[2] a being in state F carries *correlational information* about b being in state G

iff

$$P(Gb|Fa) \neq P(Gb)$$

When a carries correlational information, observing the state of a is potentially informative, at least to some extent, about the state of b. Such correlations are obviously useful, the more so the stronger they are; that is, the more a's state changes the probability of b's state.[3] An organism which needs to condition its behaviour on whether some state of the world obtains, but can't directly observe that state of the world, can instead condition its behaviour on the state of an item which carries correlational information about the relevant state of the world.

Our definition of correlational information relies on there being nomologically underpinned probabilities in the world (propensities, objective chances, nomologically based frequencies, or the like). An organism that observes a positive correlation between Fa and Gb can form an expectation, when next encountering an instance of Fa, that Gb is more likely. That expectation is well-founded if the reason for the correlation in the originally observed samples carries over to the new sample. It need not. Suppose that a particular shade of green that occurs on meat, *green-123* say, **(p.77)** is a sign of a bacterium which will multiply in the gut and lead to illness. A person could notice that eating something green-123 made them ill. They could well form the expectation that anything *green-123* is poisonous and should not be eaten. As a result, they also avoid eating some vegetables. Suppose that, in plant leaves, green-123 happens to be a sign of a toxin produced by plants to discourage herbivores. Then green-123 in a leaf does indeed raise the probability that the leaf is poisonous to eat. But it is only by accident that the correlation that exists in meat extends to plants. An organism observing the correlation in meat and projecting an expectation to plants would get things right, but only by accident. There is no nomologically underpinned correlation which explains why the expectation formed in one case should carry over to the other.

We are interested in the correlations that can be exploited by an organism to learn from samples and project to new cases, so it should be non-accidental that correlations encountered in one region (in the history of the individual or its ancestors) should project to new cases. That point generalizes. We are going to rely on the way exploitable correlations figure in causal explanations of behaviour and its success. So, we need to have recourse to correlations where it is not an accident that the correlation extends from one region to another. The correlation must subsist for a univocal reason.

Correlations can be useful if they raise probability or if they lower probability, but not if they do so unpredictably. What is useful is if there is a region where probability is raised throughout that region, or if there is a region where probability is lowered throughout that region.

Accordingly, I define exploitable correlational information as follows:[4]

Exploitable Correlational Information

Item(s) a being in state F carries *exploitable correlational information* about b being in state G

iff

(i) there are regions D and D’ such that if a is in D and b is in D’, P(Gb|Fa) > P(Gb) for a univocal reason
or
(ii) there are regions D and D’ such that if a is in D and b is in D’, P(Gb|Fa) < P(Gb) for a univocal reason

‘Region’ is intended to connote a spatiotemporal region but can be understood more widely to include sets and other kinds of collection. Items a could be all members of a species, or even all organisms, or just one individual. Where a is a particular object, the region will just be the places and times where a is (or the singleton set whose only member is a). The region could be smaller. Anya-while-adolescent may exhibit a correlation between a facial expression and a subsequent behaviour. The relevant region would then be Anya during adolescence. The items a may also be a type of object, such as human facial expressions. The restriction to regions means **(p.78)** there is no need for universality. The correlation may be highly local, such as facial expressions of Hoxton twentysomethings in the early 2010s. It is of course implicit that the items carrying exploitable correlational information are only the ones drawn from the relevant region.

The definition above is point-wise: one state raises the probability of another. In many natural cases a can be in a range of states, each of which raises the probability of b being in one of a range of other states. For example, the number of rings in a tree core correlates with the age of the tree: there being two rings makes it probable that the tree is two years old; three rings, three years old; and so on. Then F and G above can take a range of values, with each value of F mapping to a corresponding value of G about which it raises the probability.[5] An organism may learn or evolve to make use of this systematic relationship. It can then extend that expectation to new instances of the same overall relationship. A person could observe a few instances of the correlation between tree rings and age and then form the general expectation that tree age is equal to the number of rings. They may never have encountered forty-two rings in a tree core before; nevertheless, when they count forty-two rings and form the expectation that the tree is forty-two years old, that expectation is correct for an underlying univocal reason that extends from the cases they learnt about to the new case.[6] A further feature is that the different states X that a may be in exclude one another: any particular a can only be in one of these states at a time. In many cases they form

a partition, covering all the possibilities (e.g. all possible numbers of tree rings). We can define a notion of exploitable correlational information carried by a range of states as follows:

Exploitable Correlational Information Carried by a Range of States

Item(s) a being in states *X* carries *exploitable correlational information* about b being in states *Y*

iff

there are regions D and D’ such that, if a is in D and b is in D’, for a univocal reason, for every value F of *X* there is some value G of *Y* such that P(Gb|Fa) > P(Gb) or P(Gb|Fa) < P(Gb)[7]

(p.79) Animal signalling is an obvious case where correlations are exploited in the service of a function—in those cases, an evolutionary function. If there is also robustness in how the outcomes prompted by the signals are achieved, which there often is, then these cases will fit squarely within our framework. As we saw in discussing teleosemantics (§1.4), the correlations that feed into an explanation of how behaviour prompted by signals achieves its evolutionary function are correlations with distal features of the environment (e.g. with the location of nectar). Skyrms-style signalling models also turn on correlations being exploited as stand-ins on which receivers can condition behaviour (Skyrms 2010, Shea et al. 2017). (These models abstract away from the machinery of robustness.) There, which correlations are relevant can be read off directly from the payoff matrix: correlations with world states in which, given appropriate actions, payoffs are delivered.

The definition of exploitable correlational information is extremely liberal. There are very many different regions within which a correlation subsists. There will often be subregions where a correlation is stronger and larger regions where it is weaker; also partially overlapping regions. I don’t attempt to define a unique reference class with respect to which a univocally based correlation exists. There is exploitable correlational information with respect to any region within which a univocal reason extends, whether or not that reason also extends to a wider region. What counts is the region within which an organism operates: the instances of a it encounters and the instances of b on which success of its behaviour depends. The basis for the correlation is objective and independent of the organism, but the correlation strength that matters partly depends on the organism’s point of view.

For our purposes below, there has to be an exploitable correlation within the region where outputs were stabilized and robustly produced. And the correlation encountered there has to be strong enough to explain stabilization and/or robustness. What counts as strong enough will depend on the facts of the case.[8]

An extremely weak correlation might form an adaptive basis for avoiding predators, for example, because the costs of being eaten are so high. What matters in explaining stabilization is that the correlation is strong enough in the region in which stabilization occurs. If we are instead looking forward, predicting the likelihood that its behaviour will be successful, then it is correlation strength in the region where the organism will be operating that is relevant.

Correlation has been the focus of a lot of scientific work on representation in the brain.[9] At the level of individual neurons, neuroscientists have consistently looked **(p.80)** for correlations between neural firing rate and certain kinds of stimuli; for example, neurons that respond to an edge in a specific location in the visual field (Hubel and Wiesel 1962).[10] In standard region-based fMRI the search has been for regions of the brain whose activity correlates with a particular type of stimulus or task. More recent multi-voxel pattern analysis looks for correlations between the distributed pattern of activation across a region of interest and a stimulus or task type. And model-based fMRI looks for regions whose activation varies parametrically in step with quantities that the brain may be computing.

All these techniques are probing the way neural activity carries correlational information. Three features of these practices are worth noting. First, strength of correlation is always assumed to be important: carrying more information is, *ceteris paribus*, more useful; and so it is assumed that strong correlational information is a better candidate for what the brain is really representing. Secondly, the correlations being probed are very often with distal features of the environment: properties of the stimuli being presented or the task the organism is called on to perform. Thirdly, there often seems to be a tacit assumption that only information that is being used is relevant to understanding what the brain is computing (deCharms and Zador 2000). For example, there might be substantial information carried by the phase difference between neural firing rates, but that is of no interest unless there is a way for downstream neurons to detect and make use of those phase differences. That is at the level of the vehicle, and a similar constraint is often in the background at the level of the content. Incidental correlational information that just happens to be carried by a pattern of neural firing is not a candidate to figure in the computational or information-processing story unless it is somehow relevant to how the organism is behaving (Hunt et al. 2012).

(b) Toy example

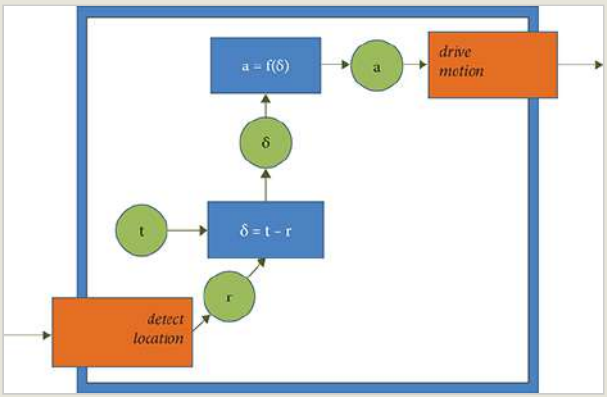
Before putting forward a concrete proposal about how correlational information gives rise to content, let’s look at a simple example in which correlation is being exploited by a system to perform a task. Consider the toy system from the last

chapter which moves along a line until it reaches a point *T*, where it stops (§3.6a).

Our toy system has four internal vehicles, **t, r, δ** and **a** (Figure 4.1). In the final version, **t** initially varies randomly across multiple episodes of behaviour, until its value is fixed by a recharge. Because of this, the value which **t** eventually adopts correlates with the location of a power source. The obvious useful correlations then are those given in Table 4.1: **(p.81)**

Table 4.1. Useful correlations carried by components of the toy system

Vehicle	Correlation
r	system's position on the line
t	location of a power source on the line
δ	distance of the system from a power source
a	velocity with which the system moves along the line



(p.83) *Figure 4.1* Toy system discussed in the text.

Those correlations make the performance of the system intelligible. That is the heart of the reason why the vehicles listed on the left of Table 4.1 are representations and the conditions on the right are their respective contents. Firing rate (let us say) of vehicle **r** correlates with distance from the origin, of vehicle **t** with location of a power source. Therefore, vehicle **δ**, whose firing rate correlates with the difference between the firing rates of **r** and **t**, will correlate with the distance of the system from a power source. The firing rate of vehicle **a** is proportional to that distance. If that rate is linearly transformed in an appropriate way into velocity, then the system will move from any point along the line so as to reach the power source. Given that those four internal elements carry the correlational information listed above, internal processing over those elements, which proceeds in virtue of vehicle properties, constitutes an algorithm for performing the distally characterized task being performed by the system (reaching T). These contents meet our desideratum (§2.2): they allow us to see why representational contents enable a better explanation of the system's behaviour than would be available without them.

(p.82) Those internal components carry lots of other correlational information—information that is less relevant to explaining how the system performs the task. For example, **r** correlates with the activity of some sensory receptors just upstream on the system's periphery. That correlation would also help explain performance, but only when supplemented with the fact that activity of the sensory receptors correlates with position along the line. So the correlation between **r** and sensory stimulation would figure in a less direct explanation of how the system performs the task. On the output side, **a** correlates with the speed of rotation of the wheels. That correlation is only explanatorily relevant because rotation of the wheels correlates with the velocity at which the system moves. So that too is less directly explanatory of how the internal components conspire together to allow the system to achieve its task function.

Suppose light falls on the engineer's workbench from one side, diminishing in intensity along the bench. Then component **r** correlates with the intensity of light at the toy's location and component **t** with light intensity at the power source. These distal correlations would also explain why the difference **δ** correlates with the distance of the system to a power source, but only when supplemented with the information that light intensity correlates with distance along the bench. So, this set of correlations carried by the internal components offers, collectively, a less direct explanation of how the system performs its task function.

We should beware of seeking determinacy where it's the wrong place to find it. And, indeed, there is some indeterminacy in what this simple toy system represents, according to my account. There are other collections of correlational information that are just as good for explaining task performance—that collectively figure in just as direct an explanation: between **t** and the location of *something worth reaching* (and between **δ** and the distance to *something worth reaching*) for example; or between **t** and the location of *a place where an outcome that reinforces behaviour and promotes persistence occurs* (with a corresponding correlation at **δ**). These alternative contents are not equivalent, since they could come apart, but they make distinctions that are more fine-grained than those that are relevant to the system. In the life of this very simple toy system being a place worth reaching is coextensional with being a charging point. Component **t** is correlated equally strongly with both. As theorists we should say that the content represented by the system is indeterminate between these options.[11] We will deal with indeterminacy in more detail in §6.2, with the benefit of the positive accounts of content set out in this chapter and the next.

4.2 Unmediated Explanatory Information

(a) Explaining task functions

In looking at a toy example in the last section, we saw that not all correlations are on a par for the purposes of explaining how a system manages to achieve its task functions. Some correlational information carried by internal components figures directly in an explanation of how a system with such internal processing

is able to perform the task and become stabilized by feedback; other correlational information is only more indirectly explanatory; and some is explanatorily irrelevant. Recall that the underlying motivation for representationalism—the practice of adverting to content properties carried by real internal components to explain behaviour—is the idea that the system's internal organization implements an algorithm for performing a task being carried out by the system. Correlations between internal elements and distal features of the environment show how a system's internal organization is keyed into the world so as to perform the distally characterized task. Content fixed in that way would meet our desideratum (to produce a theory which allows us to see how contents explain behaviour). So, the correlations that are content-constituting should be those which explain how the system achieves task functions (i.e. stabilization and robustness).

The move I make here involves a subtle shift of perspective. One could hold that content is fixed directly by its role in representational explanation: a system represents whichever contents best account for the pattern of behaviour the system produces.[12] Rather than representational explanation, my account is founded on causal explanation. Which correlations figure in causal explanations of stabilization and robustness? The former approach makes a very tight connection between what contents are and what contents explain, generating considerable indeterminacy. Causal explanations of stabilization and robustness are less indeterminate (§4.1a, §6.2).

To put this more carefully, we first define the explanandum, using a term of art, 'explaining S's performance of task functions'; then we define 'unmediated explanatory information', which is the correlational information that figures in the explanans. The explanandum has two elements, corresponding to the two elements of task function (§3.5). First, we can explain how outcomes have been stabilized (hence count as stabilized functions). Secondly, we can explain how outcomes are robustly produced (hence count as robust outcome functions). No single term is perfectly suited to encompass these two explananda. In a sense we are explaining why an outcome F is a task function of a system S, but that in turn calls for an explanation of how it was stabilized and robustly produced, so 'why' becomes somewhat misleading. 'Explaining performance of task functions' is neutral enough to cover both explananda. It also **(p.84)** emphasizes that we are focused on explaining how the system does or has done something (in its environment).

Explanandum

To explain *S's performance of task functions F_j* is to explain:

(a) how producing each of the F_j has been systematically stabilized through evolution,[13] learning or contribution to persistence (see §3.4d);
and/or
(b) how each F_j has been produced in response to a range of different inputs and achieved in a range of different relevant external conditions

Unmediated Explanatory Information

The *UE information* carried by a set of components R_i in a system S with task functions F_j

is

the exploitable correlational information carried by the R_i which plays an unmediated role in explaining, through the R_i implementing an algorithm, S's performance of task functions F_j

The idea that some correlations play an unmediated role in an explanation calls for clarification. In the classic example of the frog's fly-catching mechanism, the correlation of retinal ganglion cell firing (R) with little black things figures in *an* explanation of how the system was stabilized by evolution, but that explanation also mentions the fact that being a little black thing (condition C) correlates with being a nutritious flying object (condition C'). Without that background correlation, it would be opaque how the correlation between R and C enabled frogs to achieve an evolutionarily beneficial outcome. So, the role of the R–C correlation in that explanation is mediated. There is another explanation of stabilization that adverts directly to the correlation between R and nutritious flying objects (C'). The role of the R–C' correlation in that explanation is unmediated. A correlation between an item R and condition C plays a *mediated* role in an explanation if its role depends on the explanation adverting to a further correlation between C and some further condition C'; otherwise it plays an *unmediated* role.

The discussion in the last section effectively argued that the correlations set out in Table 1 qualify as UE information carried by the components of our toy system. (It also argued that this list is not exhaustive: there are other sets of UE information carried by the same components, hence some indeterminacy.) My claim is that, where correlation is the relevant exploitable relation, the correlational information that is content-constituting is UE information. More specifically, a sufficient condition for a vehicle to represent content p is that it carries UE information about p.

(p.85)

Condition for Content based on Correlational Information

If component R of a system S with task function or functions F_j carries UE information about condition C,

then R represents C

There is no need for an account of content to accord with what scientists relying on content think it is. For example, scientists may have no idea that content is connected to functions which are partly historically based. Nevertheless it is interesting to see that my theory of how content is constituted closely parallels a recently developed method for finding out which computations are being performed in a neural system, the method of model-based fMRI (Corrado et al. 2009). The method starts with behavioural data. For example, subjects may be asked to choose between pairs of fractal images, where different images are more or less likely to be rewarded. Subjects learn through feedback which images are rewarded when. The probabilities change during the experiment and subjects' behaviour adjusts accordingly. A large number of choices produces a rich source of data about how subjects' choices are influenced by the history of feedback they have received for past choices.

The first step is to find which computations the subjects could be performing: algorithms that are capable of producing the observed pattern of behaviour. In our terms, that is to find a list of candidate algorithms that could perform the task functions these organisms have been trained to perform. The second step is to go into the brain to see which potential algorithm is most consistent with neural activity. An algorithm calls for various quantities to be computed on the way to making a choice: expected reward, reward received at this time step, prediction error, adaptive learning rate, etc. The fMRI BOLD signal reflects the amount of neural activity in small areas of the brain, hence can reflect the quantities being represented by an algorithm implemented in the brain. We look to see if there are areas of the brain whose activity varies, trial-by-trial, with the varying quantities called for by a candidate algorithm. When areas show up as potentially representing quantities called for by the algorithm, we check that it is plausible, in terms of neural architecture, that they are computing those quantities in the right sequence. This process is repeated for other candidate algorithms and then a 'model comparison' is performed to see which algorithm is most consistent with the neural data. There are many assumptions behind the method, not all of which are strongly supported yet, but nevertheless when algorithm A fits the behavioural and neural data better than rivals B and C, that gives us some reasonable evidence that the brain is implementing algorithm A rather than B or C (Mars et al. 2012). What is striking for our purposes is that the method is effectively looking for correlational information in the brain which

explains how a person performs the task function observed in their behaviour. Model-based fMRI is looking for the properties which, according to varitel semantics, are constitutive of content.

An implementations of an algorithm has a dual character, one aspect purely local to the system and another that depends on relational properties of components of the **(p.86)** system. So, having components carrying relevant correlational information is only one half of what it takes to implement an algorithm. The other is that they should be processed in the right way—in a way that makes sense in the light of the correlational information they carry and will thereby generate appropriate behaviour. That is, when the processing is characterized in terms of local properties, independent of correlational information carried by the vehicles, it should proceed through the steps called for by the algorithm.

That in turn puts tight constraints on which correlations are likely to be explanatory, since an algorithm usually calls for different vehicles to be doing different things. For example, an algorithm might call for one vehicle that correlates with shape and another with colour, putting that information together in a third vehicle that correlates with object category. There would be *an* assignment of content to vehicles according to which all three steps simply register object category, quite noisily in the first two cases. But that set of correlational information carried by components implements a less explanatory algorithm—an algorithm that does a less good job of explaining how the system performs its task functions. For this reason, an algorithm that relies on different vehicles carrying different correlational information will generally be more suited to explaining task performance (§6.2f). We saw that at work in our toy example: an explanatory set of correlational information has **r** correlating with position and **δ** with distance to the power source (rather than both registering distance to the power source, say). In the definition of UE information, ‘unmediated’ does not count against this. To count as explanatory, an algorithm will generally have different contents at different stages. The computation of what to do is mediated through a complex series of internal states, but the job of each should be to keep track of an external condition directly, not in a way that depends on presuming a further background correlation holds.

UE information covers output correlations as well as those that are due to how a system responds to inputs. The algorithm in our toy system relies on the fact that **a** drives the toy with a certain velocity: it correlates with velocity by causing motion. Not all UE information can be about outputs, however. Part of explaining performance of a task function is to explain robustness, in particular how an output was produced in response to a range of different inputs. That will require some components of the system to carry correlational information that they reflect rather than produce. To anticipate the distinction we will discuss in Chapter 7, components can have directive contents when UE information

concerns outputs, but there has to be some descriptive content in the system somewhere.

Often output-based UE information will concern the means by which a task function is performed. In our toy example, moving with a certain velocity is a means by which the toy reaches a power source from a range of different starting positions. However, sometimes the relevant correlation will be to output an F which is itself a task function of the system. For example, humans have a learning-based task function of getting **(p.87)** sugar (in circumstances when it is needed and available). In calculating how to do that, it looks like we have an internal state in orbitofrontal cortex whose job is to correlate at input with whether we need sugar and to correlate at output with obtaining sugar (Rolls 2015, Rushworth et al. 2011, Alexander and Brown 2011). How can that output correlation be explanatory of the task function? Isn't it identical with the task function? The answer is that UE information falls out of the way the whole internal mechanism explains how outputs are produced robustly and stabilized. That requires more than just producing output F. It requires producing F in a range of different circumstances and keying it into the environmental circumstances in which it was stabilized. The algorithm as a whole is explanatory of that. A component correlating with F is one part of the overall explanation, but only when combined with other components carrying other UE information, some of which will have descriptive content. (Recall again, we are not asking which contents would best explain the behaviour; UE information is based on how an internal mechanism forms part of a causal explanation of robustness and stabilization.)[14]

This account of the way correlation can ground content is very much in the spirit of Dretske (1986, 1988). Dretske considers the case where an internal component correlates with a feature about which the system is disposed to learn, in the sense that instrumental conditioning will shape the system so as to condition its behaviour on that feature. For example, it could be because an internal state correlates with the location of a peanut (on the left or the right) that the animal comes to condition its reaching behaviour on that internal state (reaching left or right, respectively). Dretske calls the correlation with the location of the peanut a 'structuring cause' of the system's later behaviour.

That is one version of the idea that explanatory connections between correlations and the stabilization of behaviour are relevant to content determination. However, I have a more general account of why Dretske's recipe produces the right answer in the case he deals with. It is because of the role of correlations in explaining stabilization and the establishment of task functions. Instrumental conditioning of the kind Dretske points to is one specific example of that. My account is more general in three respects. First, it applies to a wider range of cases than just those which involve instrumental conditioning.[15] Secondly, my view does not require there to be pre-existing correlations **(p.88)**

between internal states and distal features. The correlations could develop at the same time as the system is being tuned to behave in a certain way. That is what happens when an artificial neural network is trained, for example. Thirdly, it applies to cases where several different correlational vehicles are involved in generating behaviour, as in the toy example we have been discussing. Dretske's recipe only applies straightforwardly where one correlational vehicle comes to be wired up to drive behaviour in virtue of the correlational information it carries.

The latter point is important, because any plausible account of representation in the brain will have to deal with the fact that very many representational vehicles interact in complex ways to produce behaviour. In §4.4 we will see a variety of ways that can arise.

(b) Reliance on explanation

The definition of unmediated explanatory information (UE information) places heavy reliance on the concept of explanation (obviously). It bases content on causal-explanatory connections. I am assuming a realist account of explanation according to which the causal-explanatory relations that figure in explanations are objective metaphysical dependence relations.[16],[17] This is not special pleading. Varitel semantics is making use of a resource here which other sciences also take for granted. It is not the task of a theory of content to give a theory of why causal-explanatory relations are objective.

Recall that contents are fixed, not by the role of contents in representational explanations, but by the role of correlations in causal explanations. So, my theory of content is not interest-relative or pragmatic. If the definition of UE information is not empty, then it picks out a property in the world. UE information then exists, irrespective of whether anyone chooses to refer to it. It might be an interest-relative matter whether we go in for explanations that appeal to this property, that is, whether we go in for representational explanations. I have been arguing that UE information (and UE structural correspondence, in the next chapter) underpins a distinctive scheme of explanation, one where correctness explains success and misrepresentation explains failure. Our epistemic interests may affect whether we appeal to this scheme of explanation.

If I'm wrong to assume that causal-explanatory relations are objective, then my accounts of content necessarily inherit any interest-relativity of causal explanation. However, the same would then be true throughout the sciences. If causal-explanatory **(p.89)** claims in all sciences are ineliminably interest-relative, then it would be no surprise that representational contents are no different.[18]

Defining a property in the way I have always raises another pressing question. It's not enough to just show that the property exists (the definition is non-empty), and exists independently of anyone's interests. Is the definition any use, does it pick out a worthwhile category? I say 'yes', of course. My argument is that UE information meets our desideratum. It allows us to explain how representational content can explain behaviour. UE information is thus a worthwhile property because, if I'm right, it is the property that figures in many explanations in cognitive science.

(c) Evidential test

Carrying UE information not only explains, but also makes it more likely that the system will achieve its task functions. That gives us another way of getting at the UE information carried by a vehicle: increasing the correlation strength of UE information should increase the likelihood of the system achieving its task functions; similarly, weakening the correlation should decrease it. So, we can supplement the constitutive condition above with an evidential test—a (fallible) way of working out what UE information is carried by elements of a system.

Evidence of UE information

For component R in a system S performing a task function or functions F_j

the correlation of the state of R with a condition G involving natural properties and objects in S's environment

whose strengthening most increases and whose weakening most decreases the likelihood of S achieving F_j

is a good candidate to be UE information carried by R

To see how this works, let's go back to our toy system. Suppose there is some random noise in the system, so that each component has a small chance of going into a random state during an episode of behaviour. Then the probability that the system is at location x, say, when **r** is in a particular state R_1, although high, is not certain. There will be some occasions when **r** is in R_1 and the system is in fact at random other locations. On those occasions, the system would not achieve its task function of reaching *T*. If the correlation between **r**'s being in state R_1 and the system's being at location x were strengthened, the system would achieve its task function more often. Similarly, weakening that correlation (increasing the noise) would reduce the frequency with which the system would reach *T*.

Now consider the correlation between **r** and light intensity. Strengthening that correlation might increase the likelihood of the system reaching *T*, provided the light intensity gradient is reasonably stable, but not by as much as strengthening the correlation with the system's location on the line would (since light intensity

is not a **(p.90)** perfect correlate of location). So, the evidential test suggests that the correlation with light intensity is a less good candidate to be UE information.

Correlations on the output side can also be assessed using the evidential test. Where slippage in the wheels or noise in the motor system impairs performance, strengthening the correlation between component **a** and the velocity of the system will have the biggest effect in increasing the likelihood that the system performs its task function of reaching *T*.

The evidential test uses the effect of correlation strength on likelihood of achieving task functions as a proxy for how directly a collection of correlations carried by components explains the achievement of those functions. However, those things are not bound to align. Nor is there a guarantee that there will be an item of correlational information that satisfies the evidential test. A correlation whose strengthening improves performance may not be such that its weakening reduces performance, for example if there is a backup mechanism that puts an effective floor on the likelihood of performing a function. Even when there is correlational information that satisfies the evidential test, that is no guarantee that carrying this information figures in an unmediated explanation of performing task functions.[19]

The evidential test is restricted to correlations with natural properties, in order to focus on correlations that are candidates to figure in a causal explanation of task performance. General principles about explanation will make complex disjunctive or gruesome properties poor candidates to figure in such explanations. (Other theorists of content have also pointed to such considerations as ruling out some problematic putative contents.) There will clearly be correlations with non-natural properties whose strengthening would do more to increase the likelihood of success. In our toy example, if the state of vehicle **r** correlated with the location of the system *and* there being no noise anywhere in the system, then success would become very much more likely. These kinds of constructed properties are much less good candidates to figure in a causal explanation of stabilization and robustness, and hence less good candidates for content.

To apply the evidential test, we need first to have a collection of candidate correlations that assigns different correlations to different vehicles. As we've just seen, that is needed if implementing the algorithm (internal processing over components) is going to explain how outputs were produced robustly and stabilized by interactions with the environment. Then we fix on one candidate correlation, hold everything else fixed, and consider what would happen if the world were changed to make that correlation stronger. Rather than being at location x 95 per cent of the time when component **r** is in state R_1, what would the effect on robustness and stabilization be if it were at location x 100 per cent

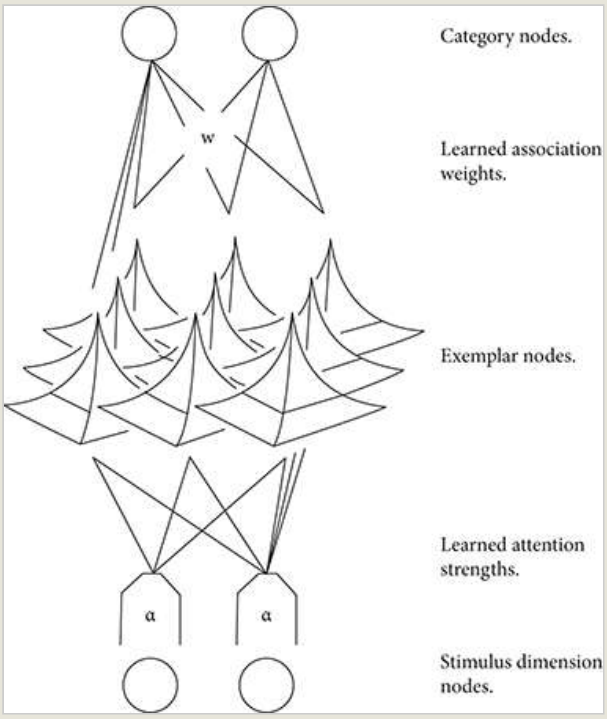
of the time when **r** is in R_1? In that case, the task-functional output (getting to the power source) would have been produced more robustly and would still have **(p.91)** been stabilized. Strengthening the correlation of **r** with patterns of sensory input would have less of an effect on successful task performance, since sensory input is itself an imperfect correlate of location. So, the evidential test suggests that **r** carries UE information about the toy's location.

It is most appropriate to assess these counterfactuals against the circumstances that obtained during episodes of stabilization. However, it is only an epistemic test. Assessing what would happen to the system in its current circumstances will also give us some evidence, to the extent that the system's current environment is relevantly similar to the circumstances in which its behaviour was stabilized.

4.3 Feedforward Hierarchical Processing

In the next six sections we see how this account of UE information can be applied to a variety of case studies. The first case is where there is simply feedforward hierarchical processing of sensory input through a series of layers. Marr's account of 3D vision is a well-known example: inputs are processed into an array of point-intensities, then into a 'primal sketch' involving detectors for blobs, edges, and so on, then on into detectors for local surfaces and their orientations, and so on (Marr 1982). Hierarchical structure is also found in the successive layers of the artificial neural networks that have used 'deep convolutional' learning algorithms so effectively to categorize a huge array of natural visual scenes (Krizhevsky et al. 2012, Kriegeskorte 2015). To work with a simpler case, consider the ALCOVE neural network model (Kruschke 1992; see Figure 4.2).

The task of ALCOVE is to categorize objects, using its sensitivity to perceptual features. To give it a clear task function, let us suppose that it has been trained to sort objects into boxes according to whether the object falls under category A or category B. The training regime gives rise to task functions because internal configurations of connection weights that tend to produce incorrect behaviour are replaced, and those which produce correct behaviour are stabilized. As a result of training, the system can use its input-sensitivity to brightness, size, etc. in order to sort objects into the correct box. Putting an object of category A into box A is then a task function of the trained system.



(p.94) *Figure 4.2* The ALCOVE network (Krushke 1992).

It performs that function by taking an intermediate step before performing the sorting action. Training produces an array of ‘exemplar nodes’ at the network’s hidden layer. These act a bit like names for individual objects. Activation of each correlates with encountering a particular object. The network solves the problem by first recognizing which individual object it is faced with, then sorting that object into the appropriate category. Input nodes correlate with features of the objects. Output nodes correlate with whether the object falls under category A or category B; they also correlate with where the object gets placed. Consider the correlational information carried by one of the exemplar nodes. Its activation raises the probability that:

(i) input nodes are activated thus-and-so
(ii) the object encountered has visual features abc (those characteristic of exemplar 1)
(p.92) (iii) the object encountered is exemplar 1
(iv) the object encountered has visual features xyz (those characteristic of objects in category A)
(v) the object encountered belongs to category A
(vi) the object encountered will be placed in box A

Those correlations are presented in decreasing order of strength ((v) and (vi) are equal). Consider how these can be combined with the correlations carried at input and output so as to implement an algorithm for sorting objects into boxes. Correlation (iii) fits together with the input correlation with object features and the output correlation with object category to instantiate an algorithm for performing the task. The correlation of hidden layer nodes with groups of perceptual features (ii) would make for a less explanatory set of correlations. An algorithm that correlated with object category (v) at the hidden layer and then again at the output layer would be less explanatory of how the system actually manages to compute what to do and perform robustly. So (iii) is UE information carried by the hidden layer. These considerations also imply that the output layer carries the UE information given by clauses (v) and (vi) above.

(p.93) How does the evidential test apply to this case? At the output layer, it is indeed the correlation with category whose strengthening and weakening most strongly affects the likelihood of success (excluding non-natural properties that would have an even tighter connection). At the input layer, it is reducing noise in the correlations with perceptual features that would have the strongest effect. Making an input node correlate more closely with exemplar or category would help in some circumstances but hinder in others, since input nodes are activated by more than one exemplar and more than one category.

The evidential test is equivocal when applied to the hidden layer. Because there is a straightforward many-to-one mapping from exemplars to categories, mistakes at the hidden layer that confuse one exemplar for another in the same category do not compromise overall performance of the system. So, tightening the connection between a hidden layer node and exemplar (iii) or category (v) will both improve performance, and to the same extent. To decide between them we have to turn to the consideration just mentioned: one collection of correlations (perceptual features, exemplar, category) provides a better understanding of how the system performs its task function than the other (perceptual features, category, category), since the latter overlooks some of the internal structure used by the system to perform the task (see also §4.1a).

Basing content on UE information has two general effects in these cases. It can select amongst coextensive properties about which the system carries information so that the most explanatory one figures in the content. And, since it is connected with explaining distal results achieved by the system, it tends to deliver contents that concern distal properties—properties that are relevant to how the system performs its tasks. So if we take JIM, a more sophisticated development of ALCOVE with more layers of processing, there is a layer of processing that detects geons—certain configurations of 3D shapes that objects can exemplify (Hummel and Biederman 1992). Does this layer represent properties of objects, or does it instead represent regular ways in which objects affect the system’s sensory apparatus? If content is fixed by UE information

carried by the layer, then it will be representing the former: properties of the distal objects.

A further development of ALCOVE uses a network with feedback connections between layers (Love et al. 2004). This raises the problem of reciprocal processing connections, which we will turn to with a different case study in §4.8 below.

It is worth noting that we have given an account of content for this system without having to give representation consumers a content-constituting role. Content comes out of how all the components interact to achieve task functions. In standard teleosemantics a consumer system is a special component, the evolutionary functions of whose outputs determine content. We noted above that it is not obvious how to extend the consumer idea to more complex cases (§1.5). That problem is not acute in this first, straightforward case study, but by eschewing a content-determining consumer even here we have an approach which is readily extendable to more complex cases.

4.4 Taxonomy of Cases

In §§ 4.5–4.8 we see how the UE approach can be applied to various cases from the empirical literature. A quick look at a typical wiring diagram for even a simple neural processing system shows that representation processing in the brain takes place in complex ways. The diagram of the primate visual system below is representative (Figure 4.3). **(p.95)**

I pick out four kinds of complexity that a theory of content will have to deal with—ways that representations are processed that show up regularly in information processing theories in cognitive neuroscience (Figure 4.4). Sections 4.5–4.8 then select case studies that exemplify each structure and show that the UE approach delivers appropriate contents. The cases are not exhaustive, but they do serve to demonstrate that the approach can be applied across a broad range of systems.

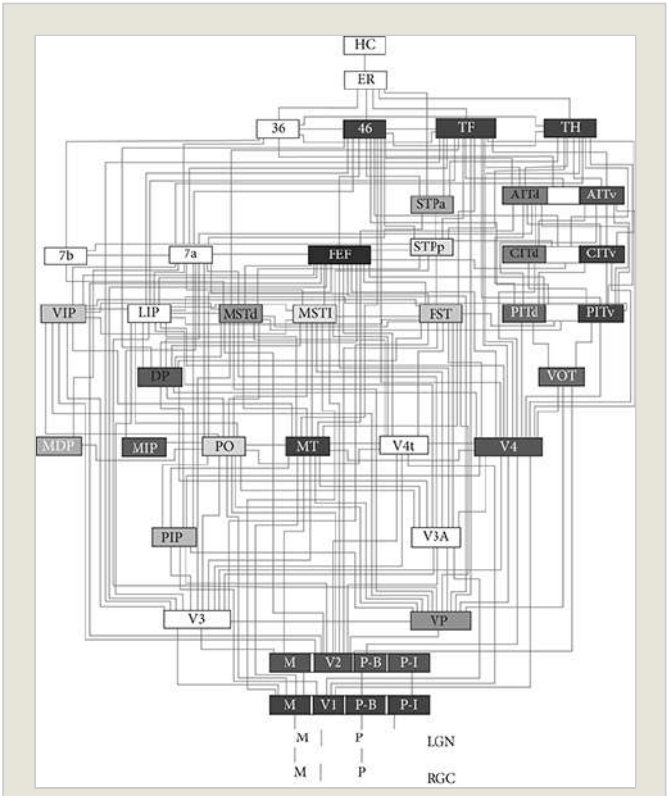


Figure 4.3 Diagram of the primate visual system (from Felleman and van Essen 1991).

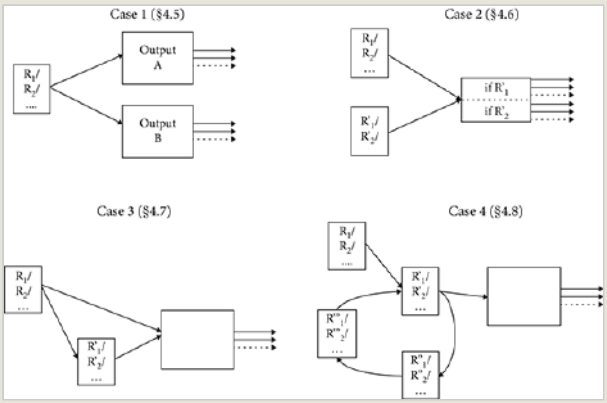


Figure 4.4 The four kinds of cases exemplified in §§ 4.5–4.8, respectively. (Inputs to the representational vehicles R and R’ are not shown.)

These cases also serve to highlight an important contrast with standard, consumer-based teleosemantics. We saw in the previous section that teleosemantics already faces a problem when we move away from cases like animal signalling where a single representational stage mediates between producer and consumer. When there are multiple layers of processing it faces a challenge in explaining why different stages represent different aspects of the problem. The consumer-based approach has an even bigger problem dealing with the kinds of cases taxonomized here, which are merely simplified treatments of some of the complex interconnections found in a typical neural processing diagram. The absence of a simple hierarchy and the presence of feedback loops make it very hard to identify, for each putative representation, a single consumer system which conditions its behaviour on that representation (Godfrey-Smith 2013; Cao 2012, 2014; cf. Artiga 2016).

The first kind of case shown in Figure 4.4 is where a single representational vehicle R, with a range of states R_i correlating with a range of properties, acts as input to two distinct subsystems conditioning their behavioural outputs on the state of R (Case 1). **(p.96)** The two subsystems may be acting for different purposes and so may be making use of different correlations carried by R. The second kind of case is the converse: two different representations are made use of by a single consumer (Case 2). For example, R may correlate with colour and also with motion, where in some contexts behaviour should be conditioned on colour and in others on motion. R’ indicates which context the system is in. In order to produce appropriate output, the organism conditions its behaviour on the conjunction of the state of R and the state of R’.

In the third and fourth cases, input influences output via more than one route. In Case 3 the two routes run in parallel. Like Case 2, the behaviour of the output subsystem depends on two different representational vehicles, but the state of the second vehicle is also dependent on the state of the first. In Case 4 we incorporate feedback: the state of vehicle R’ is affected both by current input

from R and also by input fed back from processing that took place on one of its own earlier states.

Real neural systems often include several of these elements at once. Below we pick out case studies that contain each in isolation. The aim is to show that none of them presents an obstacle to applying the UE approach.

4.5 One Vehicle for Two Purposes

A vehicle that carries correlational information about one state of affairs will usually carry information about many. Different downstream systems may be interested in different pieces of information: different correlations may be of use to each (Case 1 in Figure 4.4 above). There are many examples of that in animal signalling. A firefly produces flashes of light that signal its location to conspecifics. The signal means something like *receptive female here*. The signal also carries the information *there is a small nutritious insect here*. This second piece of information is capitalized on by predators.

In that simple case only one of the uses is part of a cooperative signalling system. Stegmann (2009, p. 873) gives an animal signalling example where both uses are at least partly cooperative. A chicken seeing a predator makes a distinctive call. This notifies conspecifics that there is a predator nearby so they can avoid it. It also notifies the predator that the chicken has seen it and could easily escape. Predator and prey share an interest in avoiding a pursuit if it will be unsuccessful. So, they share an interest in producing and acting on this signal. Conspecifics act on one piece of correlational information carried by the signal, potential predators on another.

When we turn to representations within a single organism, it is rarer that a representation is output to two entirely discrete consumer subsystems. Corollary discharge is perhaps one relatively clear case.[20] The signal sent to the motor system to drive action is also sent to perceptual mechanisms, which rely on it for the information it carries about what the animal is about to do. Very roughly, to the motor system it means *move thus and so* and to perceptual systems it means *I am actively moving thus and so*. **(p.97)** In Chapter 3 we saw that this second use of the motor signal, the efference copy, has an important role in enabling actions to be controlled smoothly.

Corollary discharge or efference copy is a very general principle of nervous systems, found even in the simplest organisms (Crapse and Sommer 2008). In more complex organisms like mammals it operates simultaneously at low levels (e.g. gating reflexes), and at higher levels (e.g. to allow predictive computations). A very simple example is found in the model organism *C. elegans* (widely studied because its nervous system has only 302 easily accessible neurons). When it senses pressure at the front, it produces a balancing motor response, which serves to stabilize its position. That reflex would get in the way when the animal pushes itself forward. So, the motor signal driving forward

locomotion also serves to cancel the stabilizing reflex, freeing the animal to make forward progress. Thus, the corollary discharge signal both instructs the animal to move forward and informs the stabilization mechanism that the animal is going to move forward under endogenous control.[21]

That is a simple case of reuse, albeit one where arguably the two contents are of different kinds: one instructs that a world state be produced (directive content) and the other is designed to reflect the world (descriptive content). On another reading there are two kinds of directive content here: telling the motor system to push the animal forward and telling the stabilizing system to be inactive on this occasion. Chapter 7 deals with the question of what makes a content descriptive or directive. The importance of corollary discharge for now is that it shows how one vehicle can have two different contents deriving from two different downstream uses.

There are also likely to be other cases where the two contents are both descriptive, as with the chicken case, rather than one being directive and the other descriptive. Section 7.4 discusses a case where that arises in a slightly subtle way. The cases we will look at below are ones where it turns out that two different systems are using the same representational content, relying on it for different (but overlapping) purposes in different contexts.

4.6 Representations Processed Differently in Different Contexts

(a) Analogue magnitude representations

One potential case of the same vehicle meaning different things in different contexts comes from the analogue magnitude system.[22] It is used to represent numerosity, but it **(p.98)** seems to be capable of representing different things in different contexts: numbers of objects, tones, light flashes, and so on. I will conclude that this is actually a case where there is a common representation of numerosity. So the case will show how representations with a common content can be processed differently in different contexts.

Analogue magnitude representations correlate with the number of objects perceived in various situations: moving visual objects, static arrays of objects, tones, taps, flashing lights, and so on. There is very good evidence showing how the analogue magnitude system works in adults, infants, and non-human animals.[23] It can be used to compare numerosity across modalities; for example, judging if the number of tones heard is more or less than the number of objects in a visually presented array. However, the correlation between the analogue magnitude register and numerosity is imperfect. The further apart two magnitudes are, the more accurate subjects will be in comparing them (5 vs. 10 is easier than 5 vs. 6), but the more things there are to compare, the less accurate the comparative judgement (5 vs. 10 is easier than 15 vs. 20). That is, the representations follow Weber’s law: discriminability is a function of the ratio of the difference between quantities to the overall quantity being compared.

There is evidence for one common register in the parietal cortex in which these numerosities are being recorded (Piazza et al. 2004, Nieder and Dehaene 2009). Registering numerosity in a common code affords ready comparisons across modalities and explains various priming and interference effects. Activation of this register R correlates with the number of items in the array or sequence presented, be they visual objects, flashes, tones, etc. Could R be a representation that has different contents for different downstream processes: numbers of objects in some contexts, numbers of tones in others, and so on? Or does R represent something common—numerosity—for all the uses to which it is put? Information about the types of item presented is not lost. Subjects can reach out to touch a flashing light, can follow moving objects with their eyes, can orient towards tones, and so on. So, there must be another component in the system somewhere with the functional role of telling downstream processing what kind of item it is dealing with, even if the number of items of that kind is recorded in a common register R. Simplifying considerably, let's suppose that contextual information about the kind of item is recorded in a separate register R'. Schematically, the set-up is the one we identified as Case 2 (Figure 4.5).

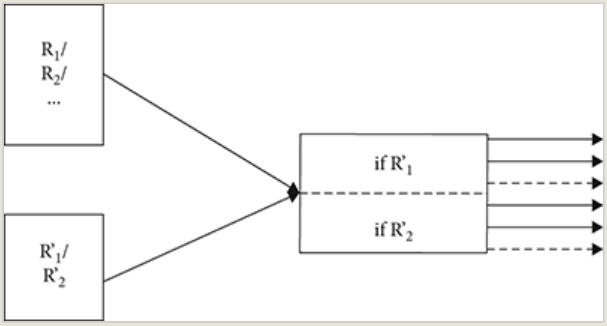


Figure 4.5 Case 2.

To home in on a task function, suppose people have been trained for monetary reward to report the number of items they have just been presented with. A visual array should be reported by pressing a button a corresponding number of times, and a sequence of tones is reported by moving a graduated slider on a screen. We can suppose that this input–output behaviour is a task function as a result of feedback-based learning. (The system will also have a more general task function as a result: to obtain money.) States R_1, R_2, etc. are states of increasing activation of register R, correlating **(p.99)** with the number of items just presented. State R'_1 correlates with the array being visual and R'_2 with it being auditory. The output behaviour is proportional to the activation of R, but the type of behaviour produced depends on whether register R' is in state R'_1, which leads to button presses, or in state R'_2, in which case the subject moves the slider.

When we look for UE information, this functional specialization is an important part of the algorithm that allows the system to perform its task functions. One register keeps track of item type and another deals with numerosity in general in a common register. That commonality is also crucial to the way the system is able to make cross-modal comparisons of numerosity. So, R comes out as carrying UE information about numerosity and R' about stimulus type. We could

treat R as representing different contents for different uses: visual objects for some downstream uses, auditory events for others, and so on. However, recognizing that R is a common register for numerosity offers a more perspicuous explanation of how the system performs its tasks than if we were to treat it has having different representations for different downstream uses. So, the UE information approach suggests that R is simply representing the numerosity of the presented array.[24] It is a common representation which combines with another representation R' to generate different outputs in different contexts.

These kinds of considerations will often be at work in applying the UE framework to real systems. Where a register is deployed in a variety of contexts, that will push in the direction of its having a common content, one which abstracts away from particular **(p.100)** sensory features of particular situations. That is, 'triangulating' on a common content, while not built into the framework, will often fall out of the explanatory considerations that underpin UE information. Perceptual representations will generally work like that. Being decoupled from any specific behavioural response also pushes in the direction of their having purely descriptive content, as we will see later (§7.4).

The analogue magnitude system also illustrates the idea that exploitable correlational information can be carried by a range of different states (§4.1a above). The activity of R varies and how active it is correlates with numerosity. That systematic relationship can extend to new cases. As a result of learning, R carries UE information according to a system that maps activity levels to numerosities: R_1 to one item, R_2 to two items, etc. Suppose that it happens that a numerosity of seven was never encountered during learning. R_7 nevertheless forms part of the same systematic relationship, so it carries the UE information that there are seven items present. Thus, when there is exploitable correlational information carried by a range of states in a systematic way, UE information can generalize beyond the instances that were encountered during stabilization (evolution, learning or contribution to persistence).[25]

(b) PFC representations of choice influenced by colour and motion

We are after a case where one and the same vehicle carries two kinds of correlational information, and intuitively one part of downstream processing makes use of it because of one kind of information that it carries, and another for another. In the previous example the functional specialization of the parietal cortex counted in favour of there being a dedicated common system for representing numerosity. So, we turn instead to the prefrontal cortex, which is less functionally specialized and carries information in a more domain-general way.

Mante et al. (2013) offer a model of information integration and context-dependent choice in the prefrontal cortex. Subjects (macaque monkeys) view an array of moving red and green dots (Figure 4.6). Sometimes the average motion is to the left, other times to the right, in each case with more or less coherence, making it more or less easy to judge the direction of coherent motion. Another dimension of variation is the proportion of dots of each colour: sometimes more red, sometimes more green. That discrimination is harder when the numbers of each colour are nearly matched. The task is either to judge the average direction of motion or to judge the preponderant colour. The task changes trial-by-trial, indicated by another stimulus (a yellow square or a blue cross at the bottom of the screen). The monkey responds by making an eye movement, either to a red circle on one side of the screen or to a green circle on the other side. When the 'colour task' stimulus is on, the monkey has to make an eye movement to the red circle if most dots are red, and to the green circle if most dots are green. When the 'motion task' stimulus is on, the monkey has to make an eye movement to the left if most dots are moving left, and to the right if most dots are moving right. **(p.101)**

Mante et al. (2013) present neural and modelling evidence that the task is performed as follows. A network of neurons in the prefrontal cortex accumulates evidence about the majority colour and the preponderant direction of motion of the dots. We can think of this simply as having a representational vehicle that varies along two dimensions, one corresponding to colour and the other to motion. The graded nature of these dimensions allows for evidence accumulation. The longer the monkey looks at a stimulus, the more information it gathers about which is the preponderant colour and direction of motion. So, the activity along these dimensions increases with time, and does so more rapidly when the difference in the array of dots (of colour or motion, respectively) is more pronounced.

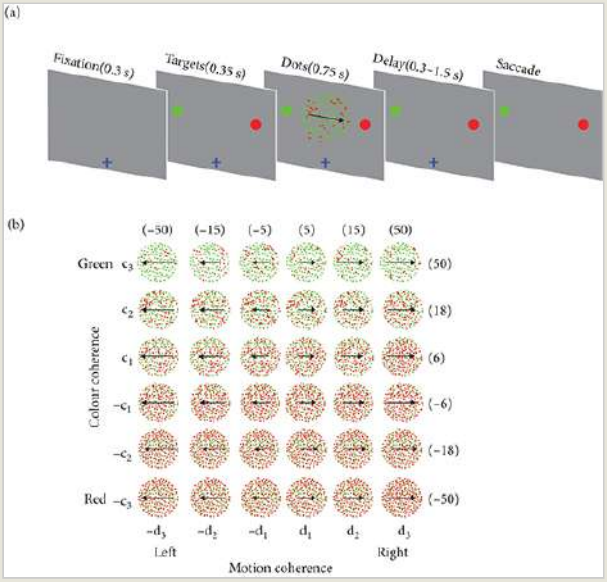


Figure 4.6 Behavioural task in Mante et al. (2013).

Activity in this network evolves over time towards one of two states, corresponding to making an eye movement in one of two directions. In the context of the motion task, evolution towards one action or the other is driven by

evidence accumulated along the motion dimension. Information along the colour dimension is preserved (indicating the proportion of dots of each colour), but it has little or no effect on which choice is **(p.102)** programmed. The reverse occurs in the context of the colour task: the colour-based dimension of variation drives evolution of the network towards a choice; motion-based information is preserved but non-efficacious. The contextual cue has the effect of selecting which dimension of variation of the representation will drive choice.

To capture the core of the case for our theorizing, let's ignore the graded activation and accumulation of evidence and consider the simplified processing diagram in Figure 4.7 (Case 2 again). We can treat the system as having two vehicles. C_1/C_2 correlates with the context cue indicating whether motion or colour will be the basis of reward in the current trial. The other vehicle can be in one of four states: R_1 or R_2 (colour) can each be paired with either R_3 or R_4 (motion). A processing step takes these states as input and produces A_1 or A_2 as output, corresponding to the two possible actions (simplifying slightly, suppose that these just program a saccade to the left or right, respectively). When in state C_1, R_1 vs. R_2 determines whether A_1 or A_2 is produced; R_3 vs. R_4 has no effect. The converse is true in state C_2.

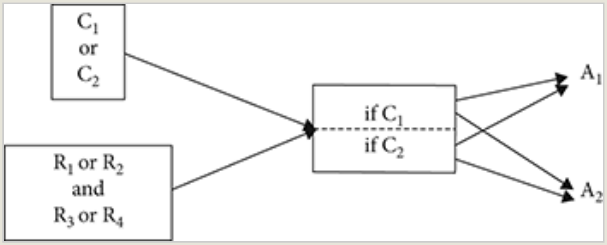


Figure 4.7 Schematic representation of the processing in Mante et al. (2013).

As a result of learning, this system has the task function of obtaining juice. To find the UE information, we need to know what worldly conditions have to obtain for the monkey to get this reward. That has been set up in this case in distal terms, in terms of properties of the stimuli. The training regime was that the left/right decision is rewarded on the basis of the preponderant colour of the stimulus in one context and the preponderant motion in the other. Amongst all the correlational information carried by the components, the correlations which directly explain achieving these rewards are:

> C_1: colour will be rewarded (if preponderant colour is red a saccade to the red circle will produce juice, if …)
> C_2: motion will be rewarded (if preponderant motion is left a saccade to the left circle will produce juice, if …)
> R_1: the preponderant colour is red
> R_2: the preponderant colour is green
> R_3: the preponderant motion is left
> R_4: the preponderant motion is right

We could consider a different set of correlational information, the correlations with sensory inputs. C_1/C_2 correlates with certain sensory states (going with the yellow square or blue cross at the bottom of the screen); states of R also correlate with activity **(p.103)** in the organism's primary sensory cortex. But an explanation of task functions in terms of this set of correlations would not be unmediated. It would have to be supplemented with further information about how neural properties relate to worldly properties. Then the correlations with the worldly properties would be doing all the explanatory work.

Lumping the states of C and R together into a single vehicle on which behaviour is conditioned would be less explanatory, for the same reasons we saw in the analogue magnitude case: it would overlook an important aspect of how internal processing manages to perform the task. Another alternative is that the system represents in a context-dependent way. In C1, it represents motion information but nothing about colour (those are mere correlations); in C2 the converse. But that overlooks the way those vehicle properties make a difference in the converse case.

Nevertheless, content attributions based on UE information retain some indeterminacy in this case. We already saw that there are two ways of capturing the UE information carried by components C: C_1 with *colour will be rewarded*, or with *if preponderant colour is red a saccade to the red circle will produce juice, if preponderant colour is green a saccade to the green circle will produce juice.* There is also indeterminacy at R, for example R_1 seems indeterminate between *the majority of the dots are red* and *the colour density in the middle of the screen is predominantly red*; or even *the moving surface in the middle of the screen is mostly red*. These collections of correlational information are equal candidates for explaining how the task of obtaining juice was performed robustly and stabilized by interactions with the environment. I would argue that finding this residual indeterminacy is the right result in this case.

4.7 One Representation Processed via Two Routes

The structure of Case 3 is illustrated again in Figure 4.8. Action is conditioned on two different representational vehicles, as in the previous section, but the second vehicle is also affected by the first. That is, the first representation affects behaviour via two routes.

Van Essen and Gallant (1994) produced an influential description of the primate visual system. One aspect of their account contains an example of the structure we are interested in (see Figure 4.9). There are several interconnections, and

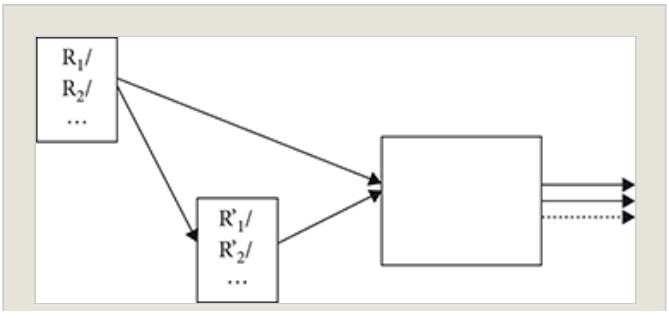


Figure 4.8 Case 3.

lots of **(p.104)** connections into and out of the visual system that are not shown in Figure 4.9.[26] We restrict our attention just to processes occurring within the visual system, and amongst those just to the stages circled in Figure 4.9. The ‘Thin stripe’ area in V2 processes wavelength information, which is input directly to MT, and also affects processing in ‘Thick stripe’ V2 which in turn also affects processing in MT.

Van Essen and Gallant were primarily concerned to catalogue the functional connections in the visual system. Their claims about what information is being processed at each stage are more tentative and are a long way from being specified computationally. In order to have a concrete example to theorize about I will simplify and then fill in some details. That will give us a simplified but more specific case to work with.

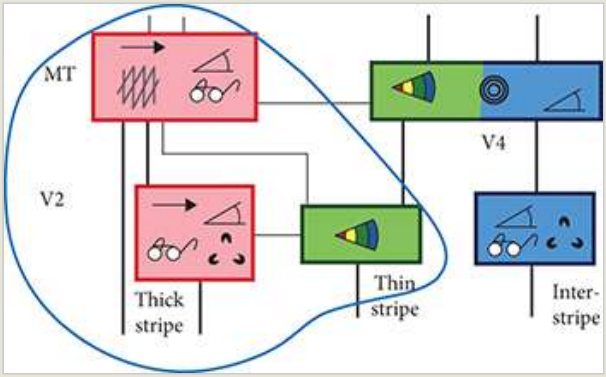


Figure 4.9 A detail from the visual information processing network described by van Essen and Gallant (1994).

Let us suppose that each box acts as a filter. Thin stripe V2 is tuned to colour dimensions, with different cells for different parts of visual space (we can think of this as retinal space). Area MT contains four different filters, but we can focus on just one, the one for plaid motion. These cells are sensitive to the direction of motion of surfaces in the visual field. They integrate local motion information and thereby correlate with the overall direction of motion of a presented surface. In some circumstances when observing gratings, that fusion can be broken so that the observer sees two superimposed gratings moving in different directions —so-called ‘plaid’ motion (Adelson and Movshon 1982; Burr 2014, pp. 766–9). We are supposing that this is in part because MT contains cells that are sensitive to the direction of motion of more than one surface in roughly the same portion of the visual field. In other conditions the observer will see just one moving grating.

This sensitivity to the direction of motion of surfaces is affected by chromatic information (Croner and Albright 1999; Thiele et al. 2001; Bell et al. 2014, p. 238). One route **(p.105)** shown by van Essen and Gallant (1994) is direct, from Thin stripe V2 direct to MT. When portions of nearby space have the same colour they are more likely to be treated as parts of the same surface. There is also an indirect effect of chromatic information, from Thin stripe V2 to Thick stripe V2

and then on to MT. Chromatic information affects the way Thick stripe V2 calculates the local direction motion. That in turn affects MT's calculation of where the surfaces are and how they are moving.

To theorize about content in this case I will simplify dramatically, so as to focus just on the dual route structure we are interested in. To give the mechanism a simple task function, suppose that the organism has been trained to reach out and intercept a moving object, doing so by tracking the direction of motion of observed surfaces. Then we can focus on one vehicle in MT, correlating with plaid motion direction, and suppose that it acts as input to the motor system so as to produce correlative reaching behaviour. That is a plausible task function if the organism has undergone training with feedback. Then the internal processing will have been tuned to enable the organism to catch objects stably and robustly.

Which set of correlational information, carried by internal vehicles, is directly explanatory of the system's ability to perform that task? The relevant MT activity correlates with the direction of motion of the surfaces of presented objects. (In our simplified setting it also correlates at output with reaching direction.) Thick stripe V2 has an array of vehicles each of which correlates with the local direction of motion of one portion of the visual field. The chromatic information in Thin stripe V2 correlates with many relevant properties; for example, with the wavelengths reflected by local areas, and with the colours of local areas. It is useful for this task because, when nearby areas have the same colour, they are likely to be part of the same surface. What is important, then, is the way activity in Thin stripe V2 for a local part of the visual field correlates with some property of presented surfaces that tends to be invariant for a given surface. Call that a chromatic-surface-property.[27]

In reality each of these components is involved in very many different task functions, and that will much more tightly constrain their contents. Even with our simplification, it is still clear that UE information will concern aspects of the distal objects the system is interacting with (e.g. motion properties), and features of the behaviour it performs on them (reaching direction). Most importantly, it is clear that the UE information will differ as between the three components we are considering. They are doing more than simply indicating *surface moving in such-and-such direction* with different levels of reliability at different stages. The problem of catching the object has been split up by having vehicles that track a series of relevant properties and performing a computation over those vehicles which is suited to calculating where to reach.

The consumer-based approach could bundle together the outputs of Thin stripe V2 and Thick stripe V2 and treat them as a single input, a vehicle which can be in a wide **(p.106)** range of states, and on which the output of MT is conditioned. The behaviour-relevant contents that can be ascribed to this

conjunctive system are simply correctness conditions like *there is an object in such-and-such region moving in such-and-such direction*. Such contents offer no insight into how the system manages to compute the integrated motion of the surfaces of objects. It throws no light on the separate roles of wavelength information and achromatically driven local motion information in performing that computation. And it entirely overlooks the dual computations performed on wavelength information in solving the problem.

In short, this is another case in which the UE information approach does a good job of elucidating the way representational explanation works—and does so without having to appeal to a content-constituting representation consumer. Varitel semantics has no difficulty dealing with cases where a representational vehicle has a dual effect on behaviour via two different routes.

4.8 Feedback and Cycles

The final case involves feedback and cyclical information processing (Bogacz 2015). Rafal Bogacz describes a fully specified computational model of how the brain calculates the probabilities used to decide between a number of available actions (Figure 4.10). The model is far from being definitively established as the truth about how the brain generates this behaviour, but it is well supported by current evidence and it suits our purposes because it is specified in enough detail to get our theorizing off the ground.

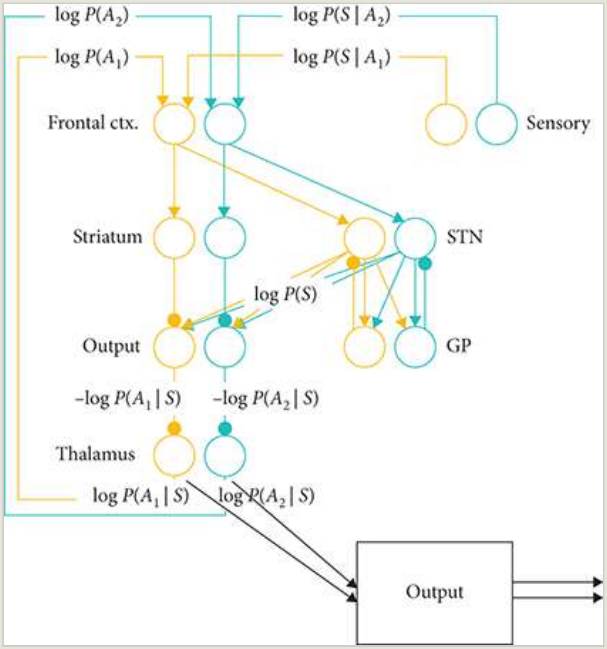


Figure 4.10 The neuronal computation proposed by Bogacz (2015) for deciding between available actions (adapted).

The computation specified in Figure 4.10 calculates the probability that each available action is the best one to perform in the current content. When one of the probabilities reaches a threshold (determined by an attempt to maximize speed at a given accuracy), that action is performed. So, let us suppose that the $P(A_i)$ are inputs to a decision layer that detects when one of the inputs crosses a threshold and programs the corresponding action (added as a rectangular box in Figure 4.10). The computation shown calls for representations, not just of states of affairs, but of probability distributions over states of affairs. That is a new feature. Before getting into the details of the

computation proposed by Bogacz, we need first to understand how the UE approach can apply to probabilistic representations.

A system can make use of the way its internal states carry probabilistic information.[28] In previous case studies we have only been concerned with the fact that a representation makes an individual world state more likely (R raises the probability that condition C obtains). But computations could make use of more fine-grained information carried by R: that when R is tokened, 75 per cent of the time the peanut is on the right and 25 per cent of the time the peanut is on the left, for example. When calculating what to do, the system can make use, not just of the fact that R raises the probability of some condition C1, but of the fact that when R is tokened, the probability that C1 obtains is p, the probability that C2 obtains is q, and so on. When this kind of fine-grained **(p.107)** correlational information figures in explaining a system's performance of its task functions, internal states will end up having probabilistic representational contents.

To apply my account to these cases, all that's needed is a straightforward extension of the definition in §4.1a of exploitable correlational information carried by a range of states. A putative vehicle of content can be in one of a range of mutually incompatible states, so it counts as a random variable X. Now consider any item in the world that can be in a range of mutually incompatible states. It is another random variable Y. There is a joint probability distribution p(x,y) relating the two states. For every possible state of the representational vehicle, this gives the corresponding probability of each of the possible states Y. One way to think about p(x,y) is just in terms of frequency: fix on one particular state of the representational vehicle and ask how often Y is then in each of its possible states; repeat for each possible vehicle state.[29] A *fine-grained exploitable correlation carried by X about Y* is a joint probability distribution p(x,y) between X and Y that subsists for a univocal reason. (In defining X and Y as random variables it is implicit that each is constituted by states of items within delimited regions D and D', respectively. We are concerned with the probability distribution that subsists within those regions.)

(p.108) The joint probability distribution thus counts as a species of exploitable correlational information. The definition of UE information applies without any modification. A representational vehicle X enters into joint probability distributions with many different conditions in the world. When a system S encounters an object, the states of X might induce a probability distribution on possible sizes, possible directions of motion, possible categories of object, whether an object is animate or inanimate, and so on. For familiar reasons, X will also induce a probability distribution on more proximal facts like states of the retina and other brain states. All these probability distributions are candidates for UE information. Which qualifies depends on which figures in explaining S's performance of its task functions. For example, the probability

distribution over motion directions may be directly relevant, given the way states of X are transformed in internal processing; probability distributions over states of the retina are less relevant.

When we were previously just concerned with probability raising, P(C|R) was important, and so was how much R changed the probability, i.e. how much P(C|R) differed from the unconditional probability P(C). That is also relevant here. Random variable X changes the probability of worldly states Y by comparison to the unconditional probabilities of states Y. This is measured by the mutual information between X and Y.[30] A set of conditions in the world about which a representational vehicle X carries more mutual information is, other things being equal, a better candidate to qualify as UE information carried by X.

We also need to generalize the idea of a correctness condition. The content represented by one of the values of X, x_1 say, is a probability distribution (call it $\hat{p}$) over world states Y: $\hat{p}(y|x_1)$. When x_1 is tokened its content would be completely accurate if $\hat{p}(y|x_1)$ exactly matched the actual probabilities of world states given x_1. When there is not an exact match we need a graded notion of accuracy/inaccuracy—of how close the content (represented distribution) is to the true distribution. There are various ways to measure how much the true distribution $p(y|x_1)$ differs from the represented distribution $\hat{p}(y|x_1)$. A standard measure is the Kullback-Leibler divergence. The Kullback-Leibler divergence of the true distribution $p(y|x_1)$ from the represented distribution $\hat{p}(y|x_1)$ measures how inefficient it is to assume that the distribution is $\hat{p}(y|x_1)$ when the true distribution is $p(y|x_1)$. It tells you how much more information (in bits) you would need in order to describe the true state of the world if you had represented it as $\hat{p}(y|x_1)$.[31] **(p.109)** This is an appropriately graded notion of inaccuracy, going to zero when the world is exactly as it is represented to be.

Returning to Bogacz's model, the computational steps are as shown in Figure 4.10. The quantities are represented on logarithmic scales so that multiplication of quantities can be performed by adding firing rates. The system starts with prior probabilities for each of the A_i. It then gets sensory input S. It can then calculate how likely action A_1 is to be rewarded given S, $P(A_1|S)$, and so on. First it calculates $P(A_1)$ $P(S|A_1)$, $P(A_2)P(S|A_2)$, etc. These then have to be normalized by dividing each by the sum of all of them so as to derive posterior probabilities for each action: $P(A_1|S)$ etc. So, the representations in frontal cortex are used in two ways. They go off to STN where they are summed, and the value of each is simultaneously preserved unmodified via the striatum, so that each separate value can be divided by this sum. If any of the $P(A_i)$ exceed the threshold at this point, the corresponding action is programmed. If not, these resulting $P(A_i)$ act as new priors for the next step of the calculation, performed on the next sensory input S.

For present purposes we are interested in the fact that information processing proceeds around a feedback loop before issuing in behaviour (Case 4, see Figure 4.11). The system has been tuned by learning to produce the action that is most likely to be rewarded given current sensory input, and to wait before acting in a way that gathers the optimal amount of sensory information to make a decision which optimizes a speed–accuracy trade off. It does so by tracking sensory information, processing it, and relying on that processing to program an action.

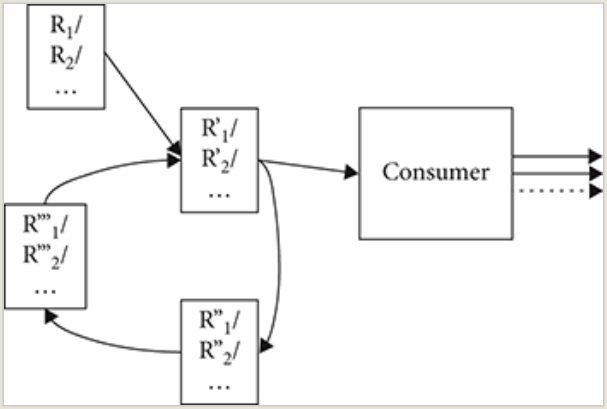


Figure 4.11 Case 4.

If the computational model above is well-supported, then it is likely to give the UE information carried by the components. According to the cognitive neuroscientists, organisms are acting near-optimally in a probabilistic environment in order to obtain the maximum amount of rewarding feedback. They look at how brain areas are probabilistically connected to states of the world in order to explain how the brain can be calculating appropriately—making calculations so that the behaviour will produce reward as often as possible. So, the test of the cognitive neuroscientific model's accuracy is the same as our test for UE information. If Bogacz (2015) is right about **(p.110)** the correlational information carried by the brain areas he points to, and if he is right that neural firing is transformed in the way he suggests,[32] then his computational model is a good hypothesis about the UE information carried by these brain areas, and hence about their representational contents.

In short, the UE information approach does allow us to give a good account of why the components of this system have the contents they do, an account which in turn feeds into an understanding of why representational contents are suited to explaining behaviour. Here again, we had no need to appeal to a consumer system that plays a special content-constituting role. Internal processing involving feedback presents no obvious obstacle to applying the varitel framework.

4.9 Conclusion

Chapter 2 argued that representational content arises in many cases from the way relational properties of components of a system combine with facts about its internal processing. Taken together, internal processing over components standing in exploitable relations to features of the environment can amount to the implementation of an algorithm, an algorithm by which the system performs

various input–output mappings. Turning this around, if we take a relevant input–output mapping, content is fixed by the exploitable relations carried by components which make the internal processing an implementation of an algorithm by which the system instantiates that mapping. Chapter 3 argued that task functions give us the input–output mappings that are relevant to content-determination. That was because of a cluster in which outcomes stabilized by natural selection, learning or contribution to persistence are also produced robustly and are generated by an algorithm that makes use of exploitable relations.

This chapter filled in the final part of the account, showing how correlational information counts as an exploitable relation within this framework. Correlations turn into content when they are exploited by a system—exploited in a very particular sense. Our definition of UE information pins down that sense: the content-constituting correlations are those which unmediatedly explain a system's performance of its task functions. We saw in this chapter how the UE approach fixes content in a range of case studies from cognitive science. It does so without having to appeal to representation consumers whose outputs play a content-constituting role. In each case study, contents fixed in this way do a good job of underpinning the characteristic explanatory grammar of representational explanation: correct representation explains successful behaviour and misrepresentation explains failure.

The next chapter argues that another exploitable relation also plays a content-constituting role, a relation in the ballpark of mirroring, isomorphism, or structural correspondence.

Notes:

([1]) We leave aside two other candidate exploitable relations because they don't arise in our simple systems: subject-predicate structure of the kind found in natural language (genuine singular and general terms); and the semantic or inferential connections between concepts, which potentially play a content-constituting role there. There may be more.

([2]) I am deliberately neutral about what should count as an item. It could be a particular object, e.g. a = the flagpole on top of Buckingham Place and F = the Union Jack is flying. Or it could be a collection of objects or a type of object, e.g. a = human faces and F = having red spots. It could also be a process or a type of process.

([3]) This 'change' need not be causal—'change' is simply convenient way of saying that the conditional probability differs from the unconditional probability.

([4]) Modifying Millikan's definition of 'soft natural information' (Millikan 2000, Appendix B).

([5]) This is closely related to Shannon’s theory of information, which connects a range of states of a receiver with a range of states of a source. Our point-wise correlational information is a special case of this. Shannon information additionally takes account of the probability distribution across this range of states.

([6]) Millikan (1984) made this important observation about evolutionary functions. Suppose that a bee dance of exactly 42.5° to the vertical has never been performed in the history of the honeybee. Nevertheless, this dance has the evolutionary function of sending consumer bees off at 42.5° to the direction of the sun. The function of the particular dance derives from the systematic relationship bees evolved to respond to (angle of dance to the vertical = direction of nectar in relation to the sun). I am making a parallel point about exploitable correlations. Where there is a univocally grounded, nomologically underpinned, systematic probability-raising relationship, expectations acquired for some values can be extended, non-accidentally, to other values drawn from the same system.

([7]) For a particular pair F and G, Fa must raise the probability of Gb across the region or lower it across the region; raising probability in some subregions and lowering it in others would not count. (This is explicit in the definition of exploitable correlational information above.) But the mapping from values of X to values of Y may be such as to raise the probability for some values of X and Y and to lower it for other values of X and Y.

([8]) There are also several aspects of correlation strength that will be important in different ways: sensitivity, specificity, positive predictive value, negative predictive value, etc. The following are always important: how likely the world state Gb is given the vehicle state Fa, i.e. P(Gb|Fa), and how informative the vehicle state is about the world state, i.e. how different P(G*b*|F*a*) is from the unconditional probability P(G*b*).

([9]) In ‘information processing’ or ‘computational’ theories in psychology and cognitive science, the ‘information’ is usually a matter of representational content rather than bare correlation.

([10]) One type of correlation is where maximum firing rate corresponds to a particular feature at a particular location, dropping off with distance or variation in the feature (e.g. rotation of a line). Another type of correlation is filtering, where it is not the maximum firing rate that is most important, but the sensitivity of changes in firing rate to changes in the stimulus. A neuron whose firing rate goes up and down substantially as the orientation of a bar changes, say, will thereby carry fine-grained information about orientation. The orientation to which it is most sensitive will be somewhere in the middle of its range of firing rates, not at the maximum.

([11]) There are different ways of capturing that indeterminacy. One is to say the representation has a correctness condition which is the disjunction of these conditions. An alternative is to say that the words we theorists use to describe the correctness condition are only an imperfect expression or model of the true correctness condition; and that each way of capturing the correctness condition using the fine-grained machinery of natural language is bound to be only approximate, each equally good.

([12]) That would make it hard to meet our desideratum.

([13]) Recall that in this case 'S' picks out a lineage of organisms, typed by a lineage-based category (§3.3). Then we have explananda like *how honeybees reach locations of nectar*, or *how* E. coli *bacteria avoid toxic chemicals*. Species (e.g. honeybee, *E. coli*) are lineage-based classes of organisms.

([14]) A different kind of case is more straightforward to deal with. Sending consumer bees off foraging 200 m away in the direction of the sun when there is nectar at that location is a task function of a bee colony. It is an output that was stabilized by evolution and robustly produced. A dance of four waggles (say) in a vertical direction correlates with sending consumer bees foraging at that location. That output correlation is with an F which is a task function. But it also explains how the colony achieves another, more general, task function: getting nectar (from a variety of different locations). So, some correlations with task-functional outputs get explanatory purchase through explaining other, related task functions.

([15]) Dretske does allow that natural selection can give an internal state the function of indicating something. The internal state can then be called a representation. But he argues, mistakenly in my view, that this is not a case where contents (reasons in his terminology) explain behaviour, since what the states indicate, 'is (and was) irrelevant to what movements they produce' (1988, p. 94), see also Dretske (1991, pp. 206–7).

([16]) This is compatible with the view that explanations are semantic entities (e.g. sentences, models); as well as with the 'ontic' view of explanation (Salmon 1984, Craver 2014).

([17]) In causal explanations, 'explains' does not introduce an intensional context, in the sense that in an intensional context it matters how we pick out the properties referred to. A neuron in a macaque's OFC that carries UE information about *orange juice* thereby carries UE information about *my favourite juice* (as it happens). Explanation does not of course in general allow substitution *salva veritate* of one property for another property which has the same extension. (In that sense causal explanations do not in general allow substitution *salva veritate* of coextensional property terms.)

([18]) As we've just seen, the relevant interests would be those related to giving causal explanations (of stabilization and robustness), rather than interests related to giving content-based explanations.

([19]) Suppose a computational step depends on comparing the values of two noisy representations and selecting the larger (as in analogue magnitude comparisons, see §4.6a below). If the noise is asymmetric around the mean, then reducing the noise in one register might cause the system to select the wrong option more often.

([20]) Thanks to Rosa Cao for suggesting this example.

([21]) Where there is an efference 'copy' it may be that there are in fact two separate representational vehicles, one instructing the animal to move forward and a separate descriptive representation telling the stabilization mechanism that the animal is going to move forward. In cases where there is just one signal, then there will be a single representational vehicle which plays both these roles.

([22]) I adopt the common label without making any claims about whether these representations qualify as analogue (rather than digital) in any useful sense; or about how best to draw the analogue–digital distinction and to characterize analogue computation.

([23]) Barth et al. (2003) in adults; Xu and Spelke (2000) in infants; Brannon and Terrace (1998) in monkeys. For reviews see Dehaene (1997) and Carey (2009, pp. 118–31).

([24]) There is a legitimate question here of what it is to represent numerosity, given that in many situations the domain being represented is discrete, whereas the vehicle of representation is either continuous-valued (e.g. a firing rate), or if it is discrete-valued (e.g. because it represents in terms of number of depolarizations, which are discrete events) then it has many more discrete values than there are integer values to be represented. Option 1 says that there are different values of the vehicle that all represent the same number of objects. Option 2 says that each state represents that the input has a certain non-integer magnitude (rational or real valued), and that it does so only approximately. How correctly or incorrectly it represents is given by the difference between the (real-valued) representational content and the (integer-valued) actual number of items; where degree of correctness can explain behavioural success and failure (e.g. the closer you are to getting it right, the more often the behaviour will be exactly appropriate to the number; and if appropriateness falls off in degrees, the more appropriate your behaviour will be).

([25]) As noted above (§4.1a), this mirrors a point made by Millikan (1984), which she describes as a kind of systematicity.

([26]) More recent work has altered this picture somewhat, but only serves to confirm that the feature we are interested in is indeed exemplified: not only is feedback vitally important, but it is also confirmed that visual processing does not occur in a strict hierarchy of subsystems or neural areas. Instead there are at least three streams that work in parallel, with interconnections between them (Shigihara and Zeki 2013).

([27]) I won't divert into considering whether these properties are identical to colour properties. In any event, there may well be some indeterminacy here: a range of surface properties that Thin stripe V2 activity correlates with, each of which is an equally good candidate for explaining performance of this particular task.

([28]) Shea (2014b) works out this idea in a canonical model of probabilistic population coding in the brain.

([29]) As before (§4.1a), this account depends on there being nomologically underpinned probabilities in the world, so these would have to be non-accidental, nomologically based frequencies; or they could be propensities or objective chances.

([30]) Taken on its own, the unconditional probability of Y is distributed across its possible states a certain way. That could be very indeterminate (e.g. all states of Y are equally likely) or already quite determinate (the unconditional probability of one or two states of Y is already quite high). This is measured by the entropy of Y, H(Y). Sharper distributions have lower entropy. States of X sharpen the distribution of Y, to a greater or lesser extent. That is to say, the entropy of the conditional distribution Y|X is less than the distribution of Y taken on its own (if X and Y are not wholly independent). This difference measures how informative X is about Y. So the mutual information between X and Y, I(X;Y) is given by the formula: $I(X;Y) = H(Y) - H(Y|X)$.

([31]) The Kullback-Leibler divergence is given by:

([32]) I.e. that neural activity in these areas is added and subtracted in the way set out in the model—a way that is suitable to implement multiplication and division of quantities that are carried, not linearly, but on a scale that is related logarithmically to activation.