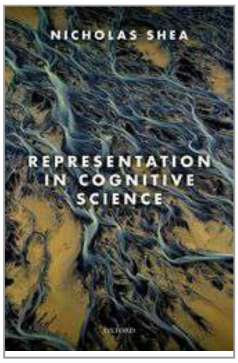

Representation in Cognitive Science
Nicholas Shea

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

Structural Correspondence

Nicholas Shea

DOI:10.1093/oso/9780198812883.003.0005

Abstract and Keywords

Structural correspondence is the other exploitable relation that figures in our case studies. It is found in the cognitive map realized by place cells in the hippocampus. When an exploitable structural correspondence is exploited in the service of a system's performance of its task functions, it is thereby constituted as a UE structural correspondence. In some cases where there is a superficially attractive structural correspondence, it can turn out that the correspondence is not being made use of; indeed, that structural representation does not arise. These cases are contrasted with two further cases where an exploitable structural correspondence is exploited. A structural correspondence may hold only approximately. That notion is defined and put to work.

Keywords: exploitable relation, correspondence, isomorphism, homomorphism, structural representation, cognitive map, UE structural correspondence, similarity judgement, causal reasoning, approximate instantiation

5.1 Introduction

Organisms and other systems make use of exploitable relations between their internal states and the world in order to perform their task functions. The previous chapter looked at correlation as an exploitable relation. This chapter argues that structural correspondence is another exploitable relation. The existence of a structural correspondence, of an appropriate kind, is part of what makes it the case that certain systems represent as they do.

Cartographic maps act as a model for theorizing about structural correspondence. Spatial relations between points on a map correspond to spatial relations between locations on the ground. Plausibly, it is because the structure of the map mirrors the structure of the world that the map gets to represent the world. A relation in one domain corresponds to a relation in the other. More carefully, a structural correspondence is a relation-preserving[1] mapping from one set of entities to another. Points on the page of an atlas map to cities, and the mapping preserves spatial relations. When point **(p.112)** a in the atlas is closer to point b than point c, then that is also true of the cities to which they correspond.

A large body of work examines the idea that structural correspondence or isomorphism should be an ingredient in a theory of content. A central problem is to specify which kinds of relation can enter into the correspondence. Just as there is a very thin notion of property, a very thin notion of relation is also available. On the thin notion of property, any arbitrary set of objects corresponds to a property. For a relation, we can capture the entities that fall under the relation with a set of ordered pairs. The relation *taller than* is captured by the set of all the ordered pairs of people where the height of the first is greater than the height of the second. On the thin notion of relation, any set of n-tuples corresponds to a relation (an n-place relation). The problem for theories of content is that the thin notion of relation makes the idea of a structure- or relation-preserving correspondence extremely undemanding. If we instead require natural relations on either side of the correspondence, it becomes too demanding; and in any event something principled must be said about why some relations should be excluded and others should count. I will use the unqualified term ‘relation’ for the thin notion, and argue for restrictions to the candidate relations on both sides of the correspondence (representation, world).

It is a familiar point, but it is useful to recall why the existence of a structure-preserving mapping or functional isomorphism is a very liberal matter. This is illustrated in Figure 5.1. There are very many such mappings between any two sets of the same size. Suppose we want a representation of relation H on a set of entities x_i. H could be the hierarchical relation of dominance between a group of macaques. Take a set of putative representational vehicles v_i of the same cardinality. For any mapping I of the v_i onto the individual macaques x_i, there is a relation V on the v_i that corresponds to H: to see if V obtains between two vehicles, map them to the corresponding individuals under I, x_i and x_j, and see if x_i is above x_j in the hierarchy (i.e. see if H obtains between x_i and x_j). That will work even if we have fixed the mapping I from vehicles to macaques (v_i to x_i) in advance.

This liberality is one of the reasons why theorists have concluded that the bare existence of a structural correspondence cannot be the basis of content (Suarez 2003, Godfrey-Smith 1996, pp. 184–7, Goodman 1972; *pace* O’Brien and Opie 2004, Cummins 1989). From our perspective the problem is that most of these correspondences are not exploitable by the system in question. Our overall desideratum is to make sense of representational explanation. We do that by content being fixed by an exploitable relation between putative representations and the world, where the obtaining of that relation explains the system’s performance of task functions. The bare existence of a structure-preserving mapping of the kind we have just seen is not something that will help a system perform a function. It is too insubstantial. So, our task is to identify a kind of structural correspondence which, when it obtains between vehicles and world, really amounts to an exploitable relation.[2]

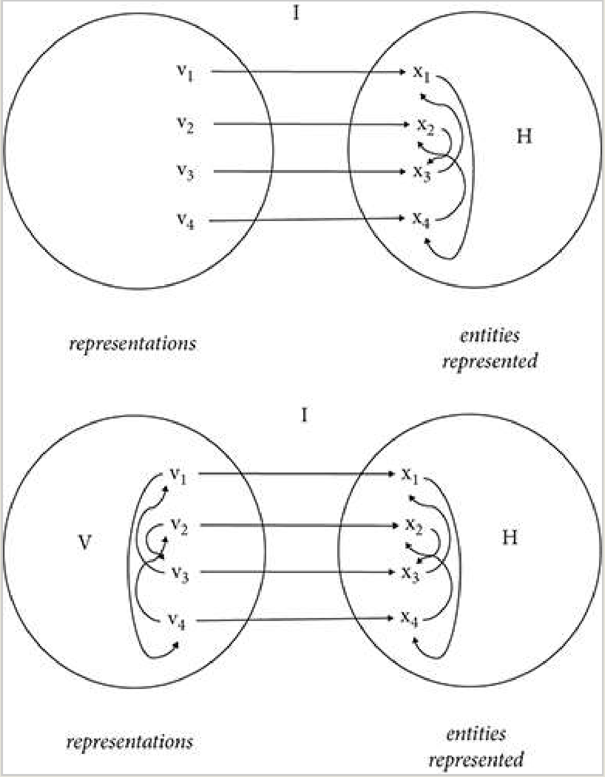

Figure 5.1 Whatever relation H we choose on the represented entities x_i (top panel), there is a corresponding relation V on the representational vehicles v_i (bottom panel).

(p.113) We start with a case study: the cognitive map in the rat hippocampus (§5.2). This shows a substantive structural correspondence in action, enabling the rat to perform task functions. After some preliminary definitions (§5.3), I narrow down this more demanding substantive sense, *exploitable structural correspondence* (§5.4a). I go on to say what it is for one of these to be exploited (§5.4b). It then counts as an *unmediated explanatory structural correspondence* (UE structural correspondence). It will turn out that, in some cases where there is a superficially attractive structural correspondence, that correspondence is not being made use of (§5.5). These cases are contrasted with two further cases where an exploitable structural correspondence is exploited (§5.6). The final section (§5.7) discusses correlation vs. correspondence, approximate correspondence, and an evidential test for telling which structural correspondence is exploited.

5.2 The Cognitive Map in the Rat Hippocampus

One of the most important neuroscientific results in recent decades has been the discovery of 'place cells' in the rat hippocampus (O'Keefe and Nadel 1978, O'Keefe and Burgess 1996). Place cells are individual neurons whose firing is specific to the **(p.114)** animal's location in space. Figure 5.2 shows the firing pattern of an array of place cells. Each panel shows the sensitivity of a single cell, with shading showing the rat's location when that cell fires, darker shading representing more vigorous firing. So, there is one cell which is active when and only when the rat is in the top-right corner of the arena (describing it from an aerial perspective); another is specific for being half-way down the left-hand side, and so on.

The array of neurons as a whole gives a very accurate measure of where the rat is in the arena. That is a useful thing to have; for example, to learn by association that certain features (smells, objects, foods) are found in certain locations (Deadwyler et al. 1996). The correlational information carried by the place cell for the top-right corner could **(p.115)** be relied on in instrumental conditioning to learn that, when at that location, the rat should pull a lever to obtain a reward. But notice that this is not to

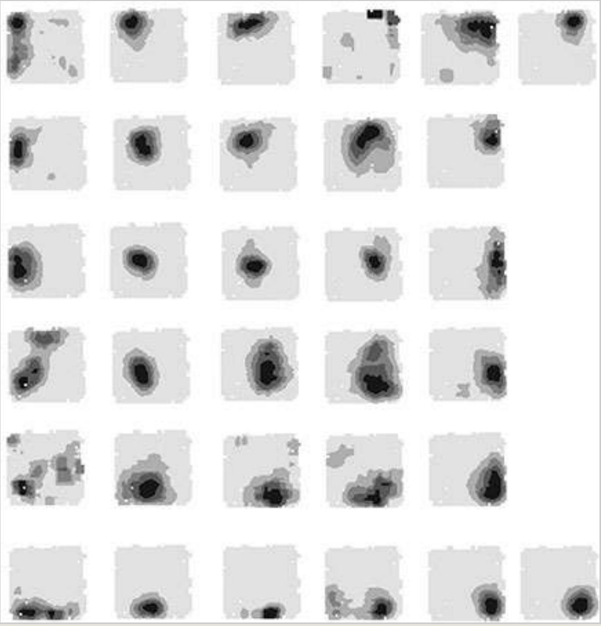

Figure 5.2 Firing rate plots for thirty-two place cells (O'Keefe and Burgess 2005). All the grey squares represent the same

square arena, each showing where a particular place cell is active as the rat moves freely around the arena. Darker shading represents higher firing rates. Different cells are tuned to different locations.

make use of a relation between the different place cells.

Indeed, taken on its own, the remarkable spatial sensitivity of the array of place cells does not depend on or give rise to any relation between the vehicles (cells). Nor are the cells spatially arranged in the hippocampus. They do not form a 'topographic map' like the retinotopic maps of visual space found in primary visual cortex. So, the remarkable discovery of the location-specific sensitivity of place cells does not, by itself, show that rats have a cognitive map. More recent work has shown however that there is an important relation on the place cells, the relation of co-activation. Cells corresponding to nearby locations tend to activate one another.

When the animal is at rest or asleep, firing of the place cells is taken offline; that is, it is no longer directly driven by input about the animal's current location (cp. stimulus-independence: Camp 2009). Offline activity shows characteristic sequences, corresponding to routes through space. 'Replay' occurs when offline sequences correspond to routes the animal has followed in the past (Wilson and McNaughton 1994, Foster and Wilson 2006, Diba and Buzsáki 2007). These connections could be built up associatively when sequences of place cells are active online as the animal explores an environment. 'Preplay' is also observed, where sequences of cells are active in advance of the animal moving, corresponding to a route the animal is about to follow (Dragoi and Tonegawa 2011, 2013). These preplay sequences lead to locations associated with reward, either because the animal has experienced rewards there in the past (Pfeiffer and Foster 2013), or because they can observe that food has been placed in that location (Ólafsdóttir et al. 2015).[3]

The current evidence suggests that rats use this prospective activity to plan the route they are about to follow. Let's suppose this allows them to select amongst possible routes, choosing a shorter one by selecting the shortest sequence of place cell firing. For simplicity, we can think of a process that activates several prospective sequences leading to a rewarded location and picks the shortest sequence as the one to follow. In fact, that search probably takes place in parallel across the whole array of place cells.[4] Either way, the co-activation structure over the place cells is being used as a proxy for **(p.116)** spatial relations between locations: this is a way of choosing an efficient route because place cells that co-activate each other correspond to locations that are close to each other in the arena.[5]

This case study fits squarely within the varitel framework. The animal has moved to a particular location *T* in the past and performed a behaviour there (e.g. pulling a lever and getting food). Its disposition to do that has been stabilized on the basis of feedback, because it received a food reward (let us say) at that location. It is then disposed and able to get to that location from a range of different starting points by a range of different routes (Pfeiffer and Foster 2013). Getting to *T* and getting food there have thereby become task functions for that individual. There is an internal-components explanation for performance of that task function: place cell activity makes use of correlational information about the current location and then proceeds offline in sequences that are driven by the co-activation structure. The animal picks an efficient route to a goal by picking the sequence that takes the shortest time to unfold during preplay. It then follows that sequence. That algorithm has been stabilized by learning in part because of a structural correspondence between co-activation on the place cells and spatial proximity on locations, relied on to calculate the route. That correspondence also feeds into an explanation of robustness, of how the animal manages to reach rewarded location *T* from a number of starting points by a range of different routes. In short, that structural correspondence is exploited. It explains the rat's performance of task functions. Therefore, I will argue, it is content-constituting: co-activation of place cells represents spatial proximity of locations.

In sum, the 'cognitive map' in the rat hippocampus illustrates how use of a structural correspondence to perform task functions can be the basis of representational content.

5.3 Preliminary Definitions

In this section I will say how I am using the terms 'structural correspondence' and 'structural representation', and what it is for a structural correspondence to play a role in constituting content. We start with structural correspondence. In all of our case studies the candidate to figure in the correspondence is some kind of relational structure. So, I define structural correspondence in terms of relations. It is a mapping under which relations are preserved.

On the world side, I will use x_i for entities and H for a relation between them. These are candidates to figure as representational contents. For example, a representation could represent that location-a is near to location-b. That would be to represent that a particular relation H obtains between two entities (locations) x_i and x_j. On the representation side, we need a way to talk about putative representations, since whether **(p.117)** they count as representations, and what they represent, flow from the obtaining of the correspondence relation. So I will call putative representations 'vehicles', v_i. V is a relation between the v_i. So the v_i potentially represent worldly entities x_i, and relation V's obtaining between v_i and v_j potentially represents that relation H obtains between x_i and

x_j. For example, the activation of place-cell-a followed by place-cell-b potentially represents that location-a is near to location-b.

The toy example of a structural correspondence in Figure 5.1 was an isomorphism, a one-to-one mapping. There are the same number of worldly entities as there are vehicles, and every worldly entity is mapped to by just one vehicle. I follow others in relying on the slightly looser notion of homomorphism. A homomorphism allows two vehicles to map to the same worldly entity. There can then be representational redundancy: two vehicles can represent the same entity. So, there may be fewer worldly entities than there are vehicles. An isomorphism is a function from some v_i to some x_i and its inverse is also a function. A homomorphism is a function from some v_i to some x_i, but its inverse need not be a function. Finally, we are interested in a homomorphism that preserves relational structure.[6] Accordingly, I define structural correspondence as follows.

> *Structural Correspondence*
>
> There is a *structural correspondence* between relation V on vehicles v_m and relation H on entities x_n
>
> iff
>
> there is a function f which maps the v_m onto the x_n and
>
> $\forall i,j\ V(v_i,v_j) \leftrightarrow H(f(v_i),f(v_j))$
>
> (*mutatis mutandis* for other polyadicities[7])

There is an issue here about structural representations and their parts. A map is a structural representation and its parts are also representations. One part of a map might consist of two points separated by 6.5 cm, representing that Cardiff is 65 km to the east of Swansea. A point taken alone can also be a representation (e.g. of Cardiff—an unsaturated representation). The definition of structural correspondence above does not require this. The vehicles taken alone need not be potential representations. This would allow for a representational icon whose parts are not themselves representations. The icon would represent in virtue of a structure over vehicles, and those vehicles would be parts of the icon, without also supposing that the individual vehicles will qualify as representations in their own right. I don’t want to take a stance on **(p.118)** whether this is possible. In all of our case studies the parts are also representations. So, I will define terms that way, following the standard definition, to stop the language becoming unbearably complex. For now, I just note that my approach could in principle apply to structures whose parts are not themselves representations.

The standard definition of structural representation does take the parts to be representations. What it takes to be a structural representation is that a relation on the representations represents a relation on the entities represented (Ramsey 2007, pp. 77–92; Swoyer 1991; Shagrir 2012). For example, spatial relations between points on a cartographic map don't just correspond to, but represent, spatial relations between the locations picked out by those points. That is a first-order resemblance, but any relation on a set of representations could in principle represent a corresponding relation on the entities represented. The obtaining of a relation between two vehicles represents the obtaining of a relation in the world. The obtaining of the relation of *being within 5 cm of* between two points on a map represents, for example, that the relation of *being within 5 km of* obtains between two cities in the world.

> *Structural Representation*
>
> A collection of representations in which a relation on representational vehicles represents a relation on the entities they represent.

We are in fact interested in something slightly stronger than there being a structural representation. Our question is whether something's being a structural representation is based on the obtaining of a structural correspondence: whether structural correspondence is partly constitutive of content. One could set up a convention in which a structure happens to represent structure in the world, but then it would be the convention rather than the existence of the correspondence which is constitutive of content. For example, I could make a list of names and stipulate that the relative size of the fonts represents the relative heights of the people named. Then there is a relation on the vehicles (relative font size) which represents a relation on the people (relative height)—so it fits the definition of structural representation—but the structural correspondence is not what is fixing content. The case studies in this chapter are stronger. The structural correspondence does fix content: the existence of a certain kind of structural correspondence is part of what makes it the case that a collection of vehicles are representations with a certain content.

> *Structural Correspondence as Content-Constituting*
>
> A structural correspondence I is *content-constituting*
>
> iff
>
> the existence of structural correspondence I, of an appropriate kind, between a relation V on putative representations v_m and a relation H on the entities x_n represented by the v_m is partly constitutive of V on v_m being a structural representation of H on x_n.

(p.119) The varitel framework applies to cases where content arises out of a system's making use of exploitable relations. So, if a structural correspondence is going to be content-constituting, it has to be used by the system. In order for it to use a structural correspondence, the system has to be sensitive in some way to the relation V between vehicles. That relation has to make a difference to downstream processing, and ultimately to the behaviour produced.

For contrast, consider the vervet monkey's system for signalling the presence of predators. Vervets make three types of alarm call for three types of danger: say R1 for aerial predators like eagles, R2 for ground predators like leopards, and R3 for snakes (Seyfarth et al. 1980). Conspecifics hearing the call make use of the fact that R1 correlates with eagles in order to behave appropriately; similarly for R2 and R3. This is a classic case of making use of correlational information. But notice that there is also a one-to-one mapping between representations and their correctness conditions (R1 to *there is an eagle*, and so on). And as ever, lots of relations are preserved by that mapping. Let's focus on just one: how high off the ground the predator is usually found. Eagles are higher up than leopards, which are higher up than snakes. I pick that arbitrarily, just to make a point. There is no evidence that the calls are telling macaques anything about height off the ground.

So, we are concerned with a relation H, *higher than*, between worldly entities. H applies to just the following ordered pairs: <eagle, leopard>, <leopard, snake>, <eagle, snake>. Now, as we've seen, there will be *a* relation on the vehicles (alarm calls) that corresponds to H. Call it V. V applies to just the following ordered pairs: <R1, R2>, <R2, R3>, <R1, R3>. So, there is a structural correspondence between relation V on the alarm calls and relation H on the predators. However, the existence of this structural correspondence is of no significance to the vervets. They are not making use of it as they process the alarm calls. They are not sensitive to whether relation V obtains between the calls. Vervets have evolved to respond appropriately to calls that have the acoustic features of R1, but that does not depend on comparing R1 to R2 or R3, or making use of any relation between R1 and the other calls. The system exploits the correlations (between R1 and eagles, R2 and leopards, and R3 and snakes), but does not exploit the structural correspondence between V and H. The structural correspondence is not content-constituting. Nor is it a case of structural representation: the relation V on the alarm calls, defined above, does not represent *higher than* (or anything at all).

The requirement that a structural correspondence be used in order to be content-constituting allows us to cut down very considerably on the problematic liberality of structural correspondence. To be content-constituting, a structural correspondence has to be exploitable. In the next section I pick out a class of structural correspondences that are candidates to be exploited. This is the restricted notion we needed: it avoids wild liberality and is restricted in a principled way. I go on in §5.4b to say what it takes for a substantive structural correspondence of this kind to be exploited, hence to constitute content.

(p.120) 5.4 Content-Constituting Structural Correspondence

(a) Exploitable structural correspondence

Recall that theories of content are faced with the problematic liberality of the general notion of structural correspondence, and hence need a more restricted notion that is substantive and well-motivated. If a structural correspondence is going to figure in varitel semantics, it has to be usable by the system, something that is a candidate to explain the system's performance of task functions. This section spells out that substantive sense. I label it 'exploitable structural correspondence'. (This is not going to be circular: it is not defined in terms of being exploitable.)

In the rat navigation case, the relation of co-activation on place cells (representations) was something that processing was sensitive to and was used in processing. That relation of course corresponds to very many relations in world, but it is the correspondence with the relation of spatial proximity on locations that makes sense of how the animal manages to perform its task functions. Spatial proximity between places is directly relevant to the task of following shortest routes to reward.

When we examine this privileged, content-constituting structural correspondence, on one end it has a relation that downstream processing is sensitive to, and on the other it has a relation in the world that is of significance to the system, given the task functions it is called on to perform. Although there being a structural correspondence is only a very weak requirement, that there should be a structural correspondence of this kind is very demanding indeed. It is a considerable achievement to have place cell activity organized in this systematic way—having a correspondence that the animal can make use of. This case exemplifies what it is for a structural correspondence to be exploitable.

> *Exploitable Structural Correspondence*
>
> An *exploitable structural correspondence* is a structural correspondence between relation V on vehicles v_m in a system S and relation H on entities x_n
>
> in which
>
> (i) V is a relation that processing in S is systematically sensitive to;
> and
> (ii) H and x_n are of significance to S.

Significance to S is significance relative to the way outcomes produced by S are stabilized and robustly produced. Sensitivity is also system-relative. Processing in rat hippocampus is sensitive to patterns of co-activation between place cells. It is not sensitive to the colour of the cell bodies of the place cells. Nor is it sensitive to where within a layer of the hippocampus the place cell happens to be located: connectivity not location of the cell is what counts. Primary visual cortex is structured retinotopically: the spatial arrangement of neurons corresponds to the spatial lay-out of retinal areas they respond to. However, the significance of the retinotopic organization is debated (Chklovskii and Koulakov 2004, Knudsen et al. 1987). A central issue in that debate is **(p.121)** precisely whether downstream processing is systematically sensitive to the spatial arrangement of neurons.

For place cells, the exploitable structural correspondence only exists because associationist learning has built up a co-activation structure. Only after that does the relation between vehicles—one place cell firing immediately after another—correspond to a relation between places which qualifies it as an exploitable structural correspondence (because only then does it have a relation on the world end, spatial proximity, which is of significance to the system). One might argue that the location-specific sensitivity of place cells is very useful even before this learning has taken place. After all, that is what allows simple associationist learning to construct a cognitive map. I don’t want to resist the idea that there is something exploitable in a broad sense even before the co-activation structure exists—the rat already has something useful. However, I use ‘exploitable structural correspondence’ in a specific sense: it requires that a relation between vehicles already exists which downstream processing is sensitive to.

There is a danger of getting confused here amongst the various relations in play. The exploitable relation is the structural correspondence. The relation of co-activation between place cells is a different relation, a relation on one side of the structural correspondence. It is not itself the exploitable relation.

Downstream processing has to be sensitive to a relation between vehicles if that relation is to form part of a structural correspondence which is exploitable by the system. Neural processing is certainly sensitive to relations between firing rates. In many cases it is also sensitive to fine-grained differences in the exact time that spikes are produced in different neurons. There are debates about whether some neural computations use a phase code, that is a code in which what counts is the time when a neuron fires in relation to a background oscillatory rhythm in the population of neurons. If so, phase differences are also candidates for the relation (V) on the representational side of an exploitable structural correspondence.

Plasticity can drive changes in the sensitivity of downstream processing. Then a relation between vehicles which previously did not count, because downstream processing was not systematically sensitive to it, may turn into a candidate. In some cases it is feedback from stabilization that drives this plasticity. In that case the exploitable structural correspondence becomes established at the same time as it contributes to stabilization. So the exploitable correspondence that is made use of to perform a task function need not pre-exist the stabilization process which grounds the task function. As just mentioned, a wider notion of exploitability is available, which does not require that the system is yet sensitive to the relation between vehicles. The category of *potentially exploitable structural correspondence* covers cases where a system can readily adjust so as to make downstream processing sensitive to a relation between vehicles, or can readily put vehicles into a relation (like co-activation) to which downstream processing is already systematically sensitive. It may well be important that some systems have access to many potentially exploitable structural correspondences. The definition **(p.122)** of exploitable structural correspondence is narrower, however, because the aim is to home in on a content-constituting correspondence. We are concerned with the actual sensitivity of the system, as configured. The class of potentially exploitable relations may in any event be less well-defined.[8] Potential exploitability certainly comes in degrees.

The definition of exploitable structural correspondence also requires that relation V should make a *systematic* difference to downstream processing. What this amounts to will depend on the types of processing in question. The general idea is that V should have downstream effects that operate according to common principles. So, when the same relation obtains between different pairs of vehicles (co-activation of two place cells), downstream processing should do the same thing in each case (treat this as a single step in calculating routes). If V comes in degrees, then processing should be systematically sensitive to those degrees. For example, if V is a difference in the time of firing, then there should be a systematic relation between the way downstream processing treats differences of 1 ms, 2 ms, and 3 ms. One way of spelling this out is to say that V should figure as a projectable property in a special science law describing the processing of the system. Whether that is the right way to understand systematic sensitivity is an issue about causation for philosophy of science more generally and not a proprietary problem for theories of content. In order not to pre-judge that issue, my definition simply makes use of the idea of systematic sensitivity, which is a resource needed throughout the sciences.

Turning to the other end of the correspondence, the things in the world being represented, the definition requires that the correspondence should be with entities x_n and a relation H that is of significance to the system. What counts as significant for the system is relative to its task functions. In the cases we are considering significance to the system narrows the candidates down to natural objects, properties, and kinds in the world. But I don’t need a general account of what naturalness amounts to: the significance requirement imports a system-relative constraint (which will require naturalness in many cases). As a result, it will mostly cut out gruesome and disjunctive properties as candidates for content, but does so in a system- or organism-relative way.

Notice that there are different constraints on the two sides of the correspondence. An obvious restriction would be to introduce a naturalness constraint on both sides of the correspondence. But any restriction needs to be well-motivated. The motivation provided by the varitel framework calls for system-relative restrictions on both sides of the correspondence but different ones. On the vehicle side, the restriction is motivated by the role of inter-vehicle relations in downstream processing. On the world side, the restriction is motivated by whether relations in the world are significant for the **(p.123)** system (significant for its performance of task functions). That is why our exploitable structural correspondence has different restrictions on each side.

On occasions when an exploitable structural correspondence is being used, relation V is instantiated between some actual vehicles, and relation H is instantiated between some actual things in the world. When I talk of a structural correspondence being instantiated, what I mean is that an instance of the relation V is instantiated between two vehicles, together with an instance of relation H being instantiated between the two worldly entities to which they correspond.

So far, we have seen the following: out of all the very many structural correspondences that exist, there are some that are ready to play a role in explaining task functions. In these cases it is a substantial achievement to have such a correspondence in place. This is the sense in which the Survey of India was such a major achievement (and such a powerful tool of colonial control). Why bother? After all, the haphazard distribution of pebbles on Horse Guards Parade already bore *a* structural correspondence to the settlements, mountains, and rivers of India (under certain mappings). What the Survey achieved was to create an artefact bearing a relation that users are readily sensitive to (spatial separation on a sheet of paper) which corresponds to a relation on the ground of significance to the colonial regime (distance). An exploitable structural correspondence is useful to have and an achievement to create. Next, we see what it is for an exploitable structural correspondence to be made use of in performing task functions: to be an ‘unmediated explanatory structural correspondence’.

(b) Unmediated explanatory structural correspondence

Our desideratum was to make sense of representational explanation, and the varitel framework achieves that by making content a matter of exploitable relations which explain performance of task functions (the explanandum spelt out in §4.2a). So, an exploitable structural correspondence constitutes content when it explains how certain outputs F_j were stabilized by one of the feedback processes in Chapter 3, and/or how they were robustly produced.

Unmediated Explanatory Structural Correspondence

A structural correspondence I between relation V on vehicles v_m in a system S performing task functions F_j, and relation H on entities x_n, is a *UE structural correspondence*

iff

(i) I is an exploitable structural correspondence; and
(ii) instantiation of I plays an unmediated role in explaining, through v_m and V implementing an algorithm, S's performance of task functions F_j[9]

(p.124) I argued at the end of the last section that the rat's algorithm for picking shortest routes exploits the structural correspondence between co-activation on place cells and relations of proximity between the places to which they are sensitive. Consider a location *T* at which the rat has previously experienced a food reward. It can get back there from a variety of starting positions by a variety of different routes, so getting to *T* is a robust outcome function. Getting to *T* is a stabilized function of the system because getting to *T* in the past (and doing something there) led to getting food, a type of feedback which reinforced the disposition to go to *T*. This outcome probably has an evolutionary function as well, deriving from the evolutionary function of the whole spatial navigation and learning mechanism. Getting to *T* and getting food there are also stabilized functions in virtue of contributing to survival of the organism. So, getting to *T* clearly meets the conditions for being a task function.

The structural correspondence between place cell co-activation and spatial relations figures in explaining how the rat gets to *T* robustly and how doing so was stabilized. Another part of the story is the UE information carried by the place cells when the rat is moving and they are online. That allows the rat to navigate from different starting points and to register when it has reached its target. Another important piece of the story is that the system carries correlational information about the location of previously encountered rewards. These correlations come together with the structural correspondence to explain how reaching *T* was stabilized by reinforcement learning. So, this is a case of UE structural correspondence. The need for convergence between UE information and UE structural correspondence is an important source of the determinacy of both kinds of content in these cases.

The final step is much shorter. UE structural correspondence is a sufficient condition for having content:

> *Condition for Content based on Structural Correspondence*
>
> If there is a UE structural correspondence between relation V on vehicles v_m and relation H on entities x_n
>
> then $V(v_i, v_j)$ represents condition $H(x_i,x_j)$

The existence of an exploitable structural correspondence is a necessary part of this sufficient condition for content. So, according to this theory, structural correspondence (of an appropriate kind) is content-constituting.

The sufficient condition for content is formulated so as to be neutral between descriptive content (*H obtains*) and directive content (*bring about H*). That distinction is discussed further in Chapter 7. The structural correspondences discussed in this chapter all underpin descriptive contents: so when the relation V is instantiated between two vehicles v_i and v_j, this represents that the relation H *obtains* between two corresponding entities. For example, when two place cells are co-activated, this represents that the corresponding locations are close to one another in space. (Which location each cell is representing is fixed by UE information.)

(p.125) My overall approach may generate a suspicion of circularity. We want exploitable structural correspondence to be a resource that can be made use of, but the terminology suggests that being usable is what makes something an exploitable structural correspondence in the first place. In fact, exploitable structural correspondence is not defined in terms of being exploitable, but in terms systematic sensitivity and significance for the system. Nor does the definition of UE structural correspondence mention exploiting a relation. So there is no definitional circle.

In most of our examples, content of the vehicles v_m has been fixed independently of the relation V over them. Names on a map represent towns and cities by convention. Rat place cells represent locations because they correlate with locations and are used to perform appropriate actions at those locations. However, there can be cases where which entities are represented and which relations are represented are both fixed in parallel. Think of a cartographic map with unlabelled points at locations (see Figure 5.3). Arguably, each point refers to a particular location. The fact that a point on the map refers to a specific location is fixed by the spatial relations of that point to other things on the map (e.g. to other points, a coordinate grid, and/or a benchmark).[10] Similarly, we could imagine introducing a new place cell into the co-activation structure of the hippocampus, but without its having any online sensitivity to location. That cell would acquire a content—would represent a location—in virtue of its co-activation relations to other place cells. So, a UE structural correspondence can determine content about entities (x_n) and their relations (H) all at once. **(p.126)**

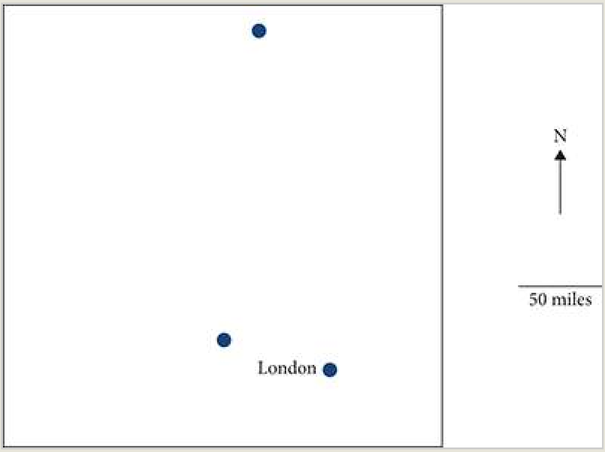

Figure 5.3 A simple map. Notice that the unlabelled points pick out locations. They do so in virtue of their spatial relations to other things on the map.

Another way a new exploitable structural correspondence can come into existence is by learning new relations on existing entities. We saw an example of that with the co-activation structure on place cells. A very different example of that is learning the sequence of count words. Taken as phonetic patterns, the count words 'one', 'two', 'three', and so on are merely arbitrarily related. There is *a* relation on them which corresponds to the mathematical relation of successor, but for the child who has not learnt to count, that relation has no significance. Learning the count sequence by rote, however, gives rise to a new relation on these phonetic patterns. Once memorized, activating one in auditory-motor imagery tends to activate the next one in the count sequence. The child then has a relation it can make use of in downstream processing.[11] This is another way that a new exploitable structural correspondence can be established over a set of putative representations, in this instance not by changing the sensitivity of downstream processing, but by altering the structures that exist over the representations.

At the personal level, mnemonics are a common way that we create structures that we can then use in reasoning. Once I've learnt a mnemonic for the first eight US presidents (will a jolly man make a jolly visitor?), I can use it to calculate temporal relations: van Buren came after Jackson and a long time after Washington. When a memorized sequence becomes automatized, like the count-word sequence, it may become possible for automatic and non-conscious processing to make use of the correspondence. And there are doubtless many cases like the rat place cells where subpersonal learning processes produce a co-activation structure that can then be used for the way it corresponds to objects and properties in the world (as in §5.6b below).

So, an organism will typically have the potential to create very many different exploitable structural correspondences, more or less easily in different cases, by constructing new relations on representational vehicles or by making downstream processing newly sensitive to an existing structure on vehicles. Such changes take a potentially exploitable structural correspondence—a category that comes in degrees and we only need to gesture at loosely—and turn it into an exploitable structural correspondence, which we have defined precisely above (§5.4a). When an exploitable structural correspondence is used to perform task functions, either when it is constructed or subsequently, it becomes a UE structural correspondence, thereby constituting content.

5.5 Unexploited Structural Correspondence

This section goes a bit deeper into the question of which structural correspondences count and which don't. In the rat navigation case an exploitable structural correspondence is exploited—it plays an unmediated role in explaining the rat's performance of **(p.127)** task functions. I start by contrasting that with a case where an obvious structural correspondence is not in fact being exploited and is not part of the basis of content.

With familiar public representations, when there is an obvious structural correspondence, we often use it. Indeed, the correspondence is often set up because of its ease of use. That is why maps use space as a representational vehicle. Many other ways of displaying data in charts and graphs also rely on space as a representational vehicle, using spatial relations to represent a very wide range of relations in the world (e.g. genetic relatedness, age, income, …).

Another common relational coding is colour. Colours are used on weather maps to represent temperature and on fMRI brain scans to represent blood flow. Relations between different regions are visible at a glance. Colours are also used in non-spatial ways of arranging data. A list of students in a class might be colour-coded by their recent test score along an axis from blue for low scores through green, yellow, and orange to red for the highest scores. Taken alone, that coding just associates a piece of correlational information with each name in the list. But the colours make it easy to compute using relations between the test scores; for example, to sort the class into three groups with similar scores, or to sort students into pairs with very different scores. These ways of using the data make use of the exploitable structural correspondence between colour space (on the representations) and relative test scores (of the individuals represented).

When the users are people, there is not much of a gap between a structural correspondence being obvious and people starting to use it. In cases in cognitive science, however, it is relatively common that an obvious structural correspondence, even one that the system could easily become sensitive to, is not exploited. As I have argued previously, the structural correspondence that exists in the honeybee nectar dance is not being exploited in my sense (Shea 2014a, pp.128–30). Although there is an obvious relation between different dances, consumer bees are not making use of relations between different dances in deciding where to fly. As standardly described, the behaviour does not take more than one dance as input. Nor does the relation between dances enter into computations in other ways. This is a case of UE information but not UE structural correspondence; indeed, it is not a case of structural representation.

There is no content-constituting structural correspondence. Condition (i) on being an exploitable structural correspondence in my sense is not met (§5.4a)—downstream computations are not sensitive to the putatively representational relation on the representational vehicles.

The bee dance has a different property which is important, and worth dwelling on briefly. There are different dances for different directions, and it is not arbitrary which dance goes with which direction. There is a systematic relation between dances which mirrors the systematic relation between directions. The system of available signs exhibits what Godfrey-Smith has called 'organization' (Godfrey-Smith 2017, p. 279). Contrast a nominal sign system like the count words for numbers. Whether a sign system counts as organized or nominal depends on what qualifies as a systematic relation between signs, and on which relations in the world are candidates to be mirrored. Does the **(p.128)** systematic relationship between signs need to be a natural relation? Does the relation mirrored also need to be a natural relation? This is similar to the question of what counts as an exploitable structural correspondence (§5.4a), but I won't attempt to resolve it here for the organized-nominal distinction.

We saw in the last chapter that often correlational information is not carried point-wise, representation-by-representation, but is carried systematically by a range of representations about a range of states (see the definition of 'exploitable correlational information carried by a range of states' in §4.1a). That is important because it allows a compact mechanism to deal with a large number of different syntactic items representing many different directions of nectar (potentially continuum-many). It extends to new cases, beyond those on which it was stabilized, when they fall into the same system. It also makes the system error-tolerant, since a representation that is incorrect but approximately true will prompt behaviour that is close to being appropriate to the situation (flying off in roughly the right direction). When UE information is based on correlational information about a range of states, the need for a systematic account that applies to a range of different representations will effectively cut down content indeterminacy. So, organization, when it exists, is an important part of the way a system of representations does its job.

Organization is sometimes assimilated to structural representation, but they are distinct phenomena. Organized signs are tokened on different occasions, during different behavioural episodes. The relation between signs is useful because the different occasions are related in a systematic way (e.g. the behaviour called for is systematically related to the direction in which nectar is located). Structural representations have parts that are tokened together during a single episode of behaviour. The structure allows the organism to behave in a way that is appropriate to the occasion. A structural representation is a single representation with representational parts; an organized sign system is a series of different representations. A structural representation must have semantically significant constituent structure; a sign in an organized sign system need not.

The parts of a structural representation need not be tokened at the same time in order to count as parts of the same representation. Parts tokened at different times can be used to calculate what to do on a single occasion. For example, place cells are activated one after the other. Their activity need not overlap in time. This is also a feature of Robert Cummins's well-known example of the driverless car (Cummins 1996, pp. 94–5; see also Ramsey 2007, pp. 198–9). The car's wheels are steered by a pin driven along a slit in a card (see Figure 5.4). When the slit is to the right of centre, the wheels are steered to the right and the car turns right (the converse for left turns). If the car is placed in a track whose turns match the card in the right way, it will follow the track without hitting the sides. Although it looks like there is a standing structural representation of the environment (the card), the way representations are tokened so as to drive behaviour is by the pin being located at different points along the card. It is the relation between these pin positions which enables the car to behave appropriately. **(p.129)**

The pin is driven along the card in a way that correlates with the movement of the vehicle along the track, forwards or backwards, at different speeds. To get a clearer view of the internal processing, imagine it unfolds step-wise, as illustrated in Figure 5.5. The car is not detecting where it is at any moment, so it needs to be started off at a location that correlates with the initial pin position. Suppose this is the start of the track. It then moves forward a certain distance. To work out how it should now align its wheels, it moves the pin forward a corresponding distance in the card and orients the wheels accordingly. This process takes two signals at input, one saying where the system is at the outset, the other correlating with how far it has moved (the turning of the cog wheels). It then makes use of spatial

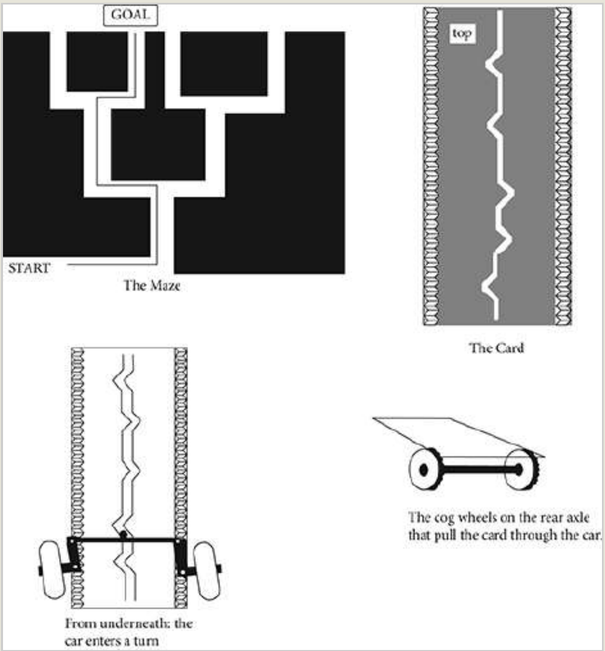
GOAL
START
The Maze
top
The Card
From underneath: the car enters a turn
The cog wheels on the rear axle that pull the card through the car.

Figure 5.4 Robert Cummins's ingenious case of a driverless car guided by a slot in a card (1996, p. 95).

relations between positions on the card to move the pin to the appropriate position on the card, and thus to act appropriately. **(p.130)**

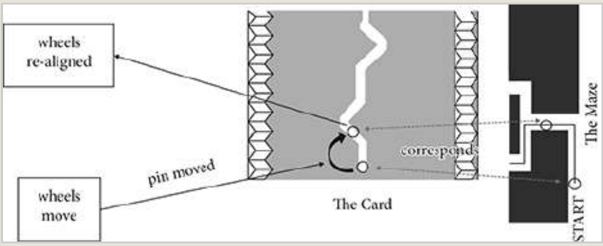

Figure 5.5 One step of the computation being performed in Cummins's driverless car case (detail).

To get to representational content we need to supplement Cummins's case somewhat, so that navigating the track is a task function of the car. We can imagine it has a task function to navigate to the end of the track as a result of robustness plus deliberate design (§3.5). We have robustness if the car is able to get to the end of the track from a range of starting positions, which would be the case if there was a mechanism to ensure that the initial position of the pin in the card correlates with the initial position of the car in the track. These additions leave the basic structure of the case intact. The car then exploits two pieces of correlational information: between the initial position of the pin and starting location; and between the rotation of the cog wheels and the distance moved along the ground. Furthermore, the mechanism exploits the structural correspondence between spatial relations on the card and spatial relations on the track. It is because spatial relations on the card correspond to distance that the system can update the pin position on the card on the basis of information about distance moved (received from the wheels). As a result of this internal computation, lengthways position of the pin on the card remains a correlate of where the car is. Widthways displacement of the pin is an instruction about how to act when at that position. Notice that if lengthways position of the pin were to correlate with location because the car is constantly detecting its current location, rather than just doing so at the outset, the structural correspondence would not be being exploited.

In his influential book *The Organization of Learning*, Randy Gallistel advances a theory of content based on isomorphism. He says he uses representation 'in its mathematical sense', which he glosses as there being what he calls a 'functioning' isomorphism between an aspect of the environment and a brain process that adapts the animal's behaviour to it (Gallistel 1990, pp. 15–33). There is a functioning isomorphism when the correspondence is exploited to solve problems in one domain using operations belonging to the other. This is clearly very like—and indeed partly inspired—my notion of UE structural correspondence. Gallistel has a further requirement: that the isomorphism should be rich, in the sense that there are many operations in the **(p.131)** representing domain that correspond to operations in the represented domain. However, in another way his requirement is much weaker than mine.

Gallistel distinguishes between direct and indirect isomorphisms. A direct isomorphism exists where the material or process embodying the representation has properties formally the same as those of the represented material or process (e.g. space mirroring space). There is an indirect isomorphism when there is no formal similarity between the representation and what is represented. For example, a mapping of mass onto written numerical symbols is only an indirect isomorphism, since ‘there is no physical ordering of the numerical symbols’ (p. 28). Gallistel allows that indirect isomorphisms, where ‘the isomorphism is created only by way of an interpretive code’, are a sufficient basis for content (p. 28).

That is too liberal, because it would apply to a downstream process that operated something like a look-up table, programming a reaction to each symbol but without relations between the symbols having any significance for the processing. A similarity in the downstream reaction is a kind of relation on the symbols, albeit indirect. (Relations on the downstream outputs other than similarity could also count.) Then there would be an ‘indirect isomorphism’ on the symbols because of the ‘interpretive code’ constituted by the downstream reactions. If we allow the interpreter and its dispositions alone to define admissible relations between representations, then we are back to the problem of arbitrary relations between representations counting. We lose the sense of the system making use of an exploitable relation. So Gallistel’s indirect isomorphisms will not in general count as cases of exploitable structural correspondence.

However, I do think there is something right in Gallistel’s idea that which isomorphisms are relevant is relative to the sensitivity of downstream processing. If processing in the rat hippocampus were not sensitive to the co-activation structure on the place cells, co-activation would not be the basis of an exploitable structural correspondence; if downstream processing then changed so that it became sensitive to relations of co-activation, the structural correspondence would become an exploitable relation. Changes to downstream processing can change which relations on vehicles are being systematically processed, but the relevant relation on vehicles cannot be a relation that exists just in virtue of similarities in the way downstream processing reacts to the vehicles. To be an exploitable structural correspondence, processing must be sensitive in some systematic way to a relation V between vehicles that exists independently of how they are used downstream. Sensitivity here is a causal notion, depending, for example, on the special science laws using projectable predicates that describe the operation of the system. That is important if there is to be a substantive sense in which the structural correspondence is a resource being used by the system. It is not entirely constituted by the way representational vehicles v_i are used.

In short, although exploitable structural correspondence depends upon the sensitivity of downstream processing, it cannot be constituted just by the way downstream processes react to vehicles. So, although exploitable structural correspondence is by **(p.132)** no means limited to Gallistel’s direct isomorphisms, it is much more limited than Gallistel’s category of indirect isomorphism.

5.6 Two More Cases of UE Structural Correspondence

(a) Similarity structure

Rat navigation gave us an example of UE structural correspondence (§5.2) and the previous section showed that seemingly obvious cases can fail to qualify. This section examines two more case studies in which a structural correspondence is exploited and is thereby constitutive of content, one involving similarity structure and the other causal structure.[12]

We can define a high-dimensional state space that captures the pattern of firing of a large population of neurons. The firing rate of each neuron in the population defines one axis in the state space. The pattern of activation distributed across the neurons at a time defines a vector in the state space. One measure of how similar two patterns of neural activity are is the distance between the two corresponding vectors in this state space (Figure 5.6). Paul Churchland is the leading proponent in philosophy of the idea that similarity in neural state space is important to the way mental representations function (Churchland 2012, 1998). Recent work analysing the distributed patterns of **(p.133)** activation recorded from neurons in non-human animals (Kiani et al. 2007) and recorded by fMRI in humans (Huth et al. 2012) has found cases where the similarity structure of neural activations does indeed mirror the similarity structure of the stimuli presented; for example, of objects of different kinds seen while watching a film.

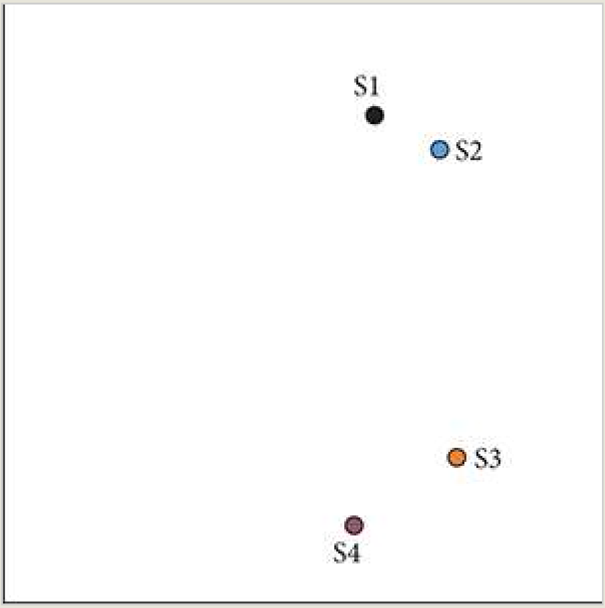

Figure 5.6 Illustration of neural similarity space. The response of two notional neurons to four stimuli S1 to S4. Responses to S1 and S2 are similar to one another and different from S3 and S4. For example, S1 and S2 could be images of faces, S3 and S4 of inanimate objects.

The existence of a similarity structure does not imply that those similarities are being used computationally, even if the similarities and differences are predictive of some observable effects like differences in reaction times, or repetition suppression in the BOLD response. However, some experiments require subjects to compute similarity; for example, if they are tasked with judging the similarity between various objects. People do so in their own idiosyncratic way. Those judgements are due in some way to how the objects are represented in the brain. Since there is good evidence that the particular structure of an individual’s similarity judgements is predicted by the idiosyncratic structure of their neural activation space (Charest et al. 2014), it is likely that similarity between neural activation patterns is the basis on which the individual is making their similarity judgements. That is, subjects are relying on a computation that uses distance in neural activation space as a measure of how similar two objects are. Another experiment used silhouettes of birds that vary along two dimensions (leg length and neck length: Constantinescu et al. 2016). When tasked with morphing an initial silhouette into a given target, subjects revealed that they had grasped the similarity space of the samples on which they had been trained. Again, this corresponded to a neural similarity space that was extracted from patterns in fMRI activation.

In line with these findings, let us suppose that activation space is sometimes used to make similarity judgements. When a subject looks at two images in series, eliciting two distributed patterns of neural activation, a measure is taken of how nearby the two patterns are in activation space. Pairs that are close on this measure are judged to be similar; pairs with a larger neural distance measure are judged to be more dissimilar. Suppose further that subjects have received feedback for correctly judging similarity according to some property of the objects.[13] Sorting objects according to similarity then becomes a stabilized function and, assuming some robustness, thereby a task function.[14] Individual patterns of activation are then being exploited for their correlation with the type of object being viewed; and the relation between two patterns in activation space is being exploited for the fact that it corresponds to the similarity of the objects represented by those two patterns. So, the correspondence between **(p. 134)** distance in neural activation space and similarity in the space of objects/properties in the world is a UE structural correspondence.[15]

These experiments raise the issue of the role of subjectively experienced similarity space: similarities and differences in the kind of conscious experience prompted by different images or objects. The experimental findings concern neural similarity space not experiential similarity space; however, a common intuition is that we use experienced similarity when judging the similarity between different objects. That is not the claim made here. My claim that relations between patterns of neural activation can structurally represent similarity between objects does not depend on these similarities and differences being experienced by the subject. The way content arises out of relations between vehicles does not depend on those relations being apparent at the personal level.

(b) Causal structure

The second case involves causal structure. The cognitive details are less clear, but the case is important because the ability to represent causal structure has been so significant for the evolution of human cognition. It is through understanding causal structure that we are able to assess the effects of various interventions. For example, we can observe that a falling barometer needle predicts that it will rain but, understanding the causal structure, we wouldn't try to make it rain by moving the barometer needle. Causal understanding is crucial to human tool use and technology.

Many animals can learn which action is best to perform in a situation. A simple way to do that is to keep track of the consequences of performing each action, and to value an action more when it produces good consequences. That way of learning does not record what the consequences were, just whether they were good or bad. It is called 'model-free' or 'habit-based' learning. It does not involve a causal model of how actions produce their consequences. The animal gets into the habit of performing an action when it has repeatedly led to good consequences. Action A could get a high value because it leads to water and happened to have been performed when thirsty. If the animal is no longer thirsty, then getting water is no longer rewarding, but action A would still be chosen. It takes a number of trials to learn that action A now no longer leads to rewarding consequences. A system with knowledge of causal structure, by contrast, can represent that action A leads to water. This allows a person to calculate, when they are not thirsty, that the consequences of performing action A are no longer valuable. They can refrain from choosing it without having to experience the consequences. Decisions based on reasoning with a causal model of actions and their consequences are called 'model-based' or 'goal-directed' (Dayan 2014). The habitual tendencies **(p.135)** produced by the model-free system can be inhibited to allow the person to choose a model-based or goal-directed response.

A now-classic way to test for model-based reasoning, hence for knowledge of causal structure, is the two-step task (Gläscher et al. 2010). This adds probability into the picture. Suppose you are presented with sweets wrapped in black or white wrappers, one colour for strawberry, the other for lemon, and you don't know which is which. The sweets are in two jars, jar A has mostly black sweets, jar B mostly white. You like lemon and hate strawberry. You reach into jar A, which is mostly black, but happen to get a white sweet, and find that it is lemon-flavoured. Your action, reaching into jar A, was rewarded. So, the model-free system would incline you to do it again. Instead you reason that you are more likely to get the lemon flavour you want from jar B, because white sweets are much more numerous there. So you do the opposite of your previously rewarded action and reach into jar B. Experiments with this logic show that human subjects do select actions based on knowledge of causal structure (Gläscher et al. 2010, Daw et al. 2011). However, we have not yet reached structural representation, because the computations involved in this reasoning only require correlation-based representations of states and of transition probabilities between states (Daw and Dayan 2014).

A more complicated experiment does give us evidence that humans have structural representations of causal structure. Quentin Huys and colleagues trained subjects on the task structure illustrated in Figure 5.7 (Huys et al. 2012, Huys et al. 2015). Think of making a series of left–right choices as you move through a maze. Subjects had to make a series of three to five binary choices to pass between six boxes, with the cost or benefit of each choice dependent on which box the subject was in when choosing. For example, when in box 1 a left button press produces a reward of 140 pence and a right button press a reward of 20 pence. Subjects never see the structure of the task but have to learn it by making a series of choices and getting feedback.[16] Huys et al. were able to test rival models of which calculations were driving subjects' behaviour and obtained good evidence that subjects were indeed evaluating in advance the overall benefit of possible sequences of choices before making their decisions. These calculations involve partial searches and maladaptive 'pruning': subjects overlook optimal sequences if they involve a large initial loss.

Causal planning is likely to depend on representations in the prefrontal cortex, especially when a hierarchy of steps is involved (Koechlin et al. 2003; Passingham 2008, pp. 168–70; Koechlin and Hyafil 2007; Balaguer et al. 2016). Understanding how a series of actions and events are causally linked may be an elaboration of the ability to represent the sequential order of events. We saw above that the rat hippocampus will replay activity corresponding to a sequence of locations the animal has visited. **(p.136)** Similarly in a non-spatial task, when human subjects learn sequences of six visual images, brain activity during rest spontaneously revisits the states it was in when viewing the images, capturing the order in which the images were experienced (Kurth-Nelson et al. 2016).[17] When sequential structure mirrors causal structure, that correspondence is exploitable for the purposes of causal reasoning.

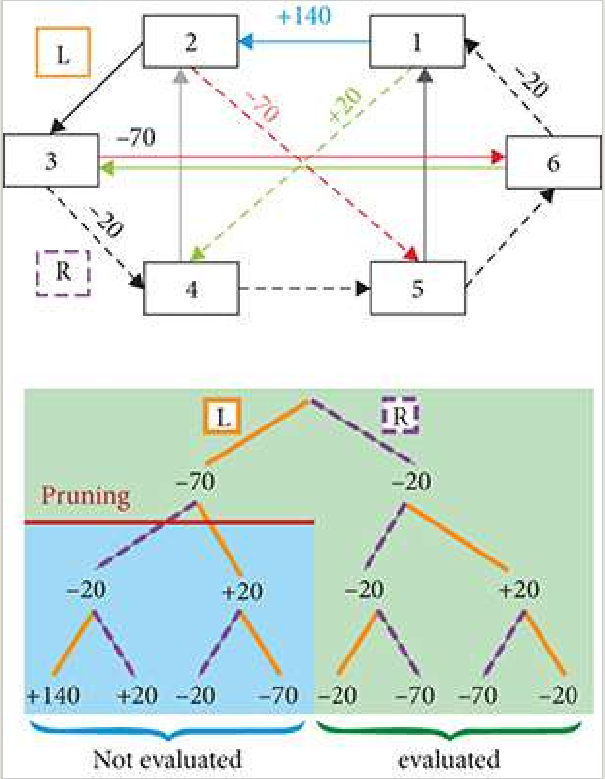

Figure 5.7 The top panel shows the structure of the task studied by Huys et al. (2012, 2015). Arrows starting at each box are labelled with the reward or cost of choosing the left option (solid arrow) or right option (broken arrow) at that step. E.g. choosing right in box 1 produces a small gain of 20 pence and leads to box 4. The bottom panel shows one of the decision trees which subjects think through when they are evaluating possible paths through the structure starting from box 3. Subjects don’t evaluate the left-hand branch (solid line) beyond the first step, since it involves incurring a large initial loss (−70 pence), even though it would be the optimal choice (left, right, left for +50 pence in aggregate).

The calculations ingeniously uncovered by Huys et al. (2015) clearly depend on subjects representing the relations between the six states, reasoning through sequences of them, and sometimes cutting that reasoning short when they encounter a large loss. There is not a rich neural story of how the step-by-step reasoning occurs, but the findings of Kurth-Nelson et al. (2016) are suggestive. So, let us suppose that subjects have brain states that occur in sequential order;

for example, the state for box 1 potentiates the states for box 2 and box 4, each conditional on a different action (left and right, respectively). When a subject calculates that box 5 is accessible within two steps from **(p.137)** box 1, that calculation makes use of the sequential structure of brain states, and of the fact that it corresponds to causal structure in the world in which she is making her choices. That would then be a case of UE structural correspondence. The sequential order of neural states is being exploited for its correspondence with the relation of causal accessibility between world states. In the absence of detailed understanding of the neural vehicles, this is more of a 'how-possible' case study. It does show how UE structural correspondence could be a suitable resource to form the basis of structural representations of causal structure.

5.7 Some Further Issues

(a) Exploiting structural correspondence cannot be assimilated to exploiting correlation

An objection to basing content on UE structural correspondence runs as follows: any exploitable structural correspondence carries correlational information and in fact it is the correlational information that is playing the content-constituting role. I agree that in very many cases the relation V involved in a UE structural correspondence will also carry correlational information about the relation H it represents. The relation of co-activation between place cells is learnt. It's being instantiated raises the probability that the two corresponding locations are near to one another. Even if a structure is acquired by evolution, and is not subject to learning during the lifetime of an individual organism, there is still often a sense in which it carries correlational information: had the world been different, the structure would have been different. So, the structure being as it is raises the probability that various relations obtain in the world.

However, the fact that a relation V between representations v_i and v_j carries correlational information does not imply that V's carrying information is being exploited, nor further that it is being exploited for carrying information about the obtaining of a relation *between the entities represented by v_i and v_j*. Think about hierarchical processing; for example, Marr's theory of the stages of processing in the visual system (Marr 1982). Activity at one layer in the hierarchy depends on the activities of vehicles lower down the hierarchy, in particular on relations between them. For a simplified example, consider the way angular disparity between the two eyes is used as a depth cue (see Figure 5.8). When the eyes focus on an object, the more their viewing angles converge, the closer the object is. Various signals in the brain correlate with eye gaze direction: let's suppose state A is a firing rate that correlates with and represents the horizontal angle of the left eye, state B of the right. The difference between rate A and rate B correlates inversely with the distance of the object of focal attention. That is, a relation between A and B, call it C, correlates with the distance of the object. Suppose downstream processing makes use of this relation C in a way that depends on the distance of the object;

for example, by programming reaching movements that correlate with C. Is C thereby a structural representation? **(p.138)**

To qualify as a structural representation, the relation C on vehicles A and B would have to represent a relation on the entities represented by A and B (see definition in §5.2 above). That is not the case here. The content of C is something like *the attended object is at distance x*. A and B represent eye direction (e.g. something like *the left eye is pointing at angle θ*). C is not representing a relation between the entities that figure in the contents carried by A and B. Hierarchical processing will make use of relations between representations to extract further useful information from them. That involves coming to represent a new condition that could be inferred probabilistically from the conditions already represented. It does not generally involve representing a relation between entities already represented.

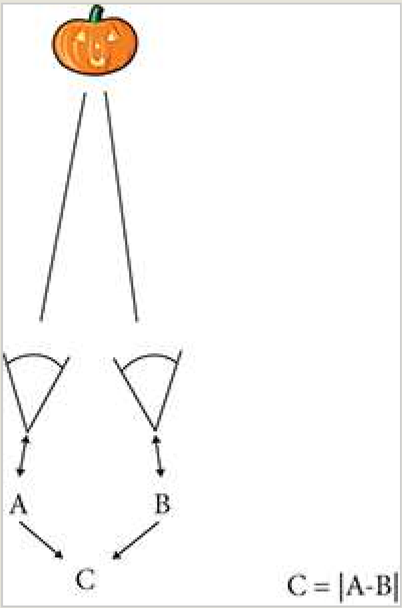

Figure 5.8 Firing rates A and B correlate with the current orientation of the left and right eyes, respectively. The (unsigned) difference in firing rates C correlates inversely with distance to the current object of focal attention.

A different line of objection is that my account of UE information already trades on a second-order resemblance theory of content. I have a collection of internal vehicles performing a computation. The functional relations between vehicles seem to correspond to relations between the entities they represent. For example, vehicles representing local motion and local colour are transformed into a vehicle representing coherent motion (§4.7). That functional transition seems to correspond to a relation in the world: surfaces exhibiting such-and-such local chromatic patterns tend to be moving thus-and-so. Doesn't the whole story about internal vehicles implementing algorithms depend on functional resemblance fixing content; that is, on a second-order correspondence at the level of computational structure (see O'Brien and Opie 2004)? The answer is that a computational step is not itself a structural representation. It does not represent that a relation obtains in the world. The utility of a computational step might depend on the presupposition that p (e.g. that such-and-such chromatic properties are a sign of sameness of surface). We could even say that the system implicitly represents that p (Shea 2015). But this is not a content for which there is a vehicle. The information **(p.139)** that p obtains is not available to be calculated with, to be used in other computational steps. You can call this a 'computational structure' if you like, but that does not entail that structural representations are involved.

So, most cases where a relation between representations is exploited for its correlational information, and therefore carries UE information, do not qualify as cases of structural representation. Exploiting structural correspondence is a special kind of case, which makes it worthwhile to pick it out and analyse it separately. And, indeed, the way content is constituted works differently.[18] That has two consequences. The first consequence we saw with place cells: a new place cell would have content in virtue of its place in the co-activation structure, irrespective of any online correlational properties. With structural representations based on UE structural correspondence, since the same relation has a systematic significance across a range of representational vehicles, new representational vehicles that fall under the relation can acquire content in a way that is independent of their correlational properties. The second consequence is exemplified by the way co-activation is used to calculate efficient routes: the relation is available to be used in computations across a range of vehicles in a systematic way. Neither of these features necessarily accompanies UE information.

Furthermore, it is at least conceptually possible for there to be a UE structural correspondence that carries no correlational information at all. An ant crawling in the sand could by chance trace a figure that looks like Winston Churchill (Putnam 1981, p. 1). The sand figure would carry no correlational information, but someone noticing the structural correspondence could use the figure to make calculations (e.g. comparing eye separation to nose length). Similarly, in subpersonal cases, a structure that happened just by chance to correspond in useful ways to significant entities and properties in the world would be useful to an organism, even though the structure's being that way is accidental—it carries no information about relations in the world of significance to the organism. It is not so far-fetched that there could be accidental structural correspondences that neural computations can make use of. Neural activity can organize spontaneously into cycles, automatically proceeding through a repeating series of steps.[19] Such a cycle bears a structural correspondence to all kinds of cyclical processes in the world (recall the liberalism) without carrying information about them. For example, a ten-stage neural cycle corresponds to ten major stages in the life cycle of a perennial plant. Neural processing can readily become sensitive to the time that it takes to transition between states of a rapid neural cycle. Then temporal relations **(p.140)** between stages in the plant cycle could be computed by using the (much shorter) temporal relations between states in the neural cycle. In this way a purely accidental correspondence would come to be a UE structural correspondence.

In sum, there are good reasons for a theory of content to pick out UE structural correspondence separately from UE information as a basis for the existence of representational content.

(b) Approximate instantiation

The definition of structural correspondence we have been using is an exact one. That is an idealization. The way that a structural correspondence explains task functions is that instances of it are instantiated[20] on occasions when the task function is stabilized and robustly produced. The correspondence does not need to be exact on those occasions in order for the structural correspondence to be explanatory. (Similarly, a correlation does not need to be perfect for correlational information to be explanatory.) A correspondence with two locations being roughly 10cm apart during stabilization can explain an organism's performance of task functions.

Consider a structural correspondence I that maps co-activation to distance with a certain metric, and consider a certain co-activation delay V that occurs between the activity of two place cells v_i and v_j. I maps these to locations x_i and x_j and maps V to a spatial separation of 1 cm. We can say I is *approximately instantiated* on occasions when the actual distance between x_i and x_j is approximately equal to the distance to which V maps under I, in this case 1 cm. The explanandum is near-optimal behaviour, and the fact that I is approximately instantiated can explain why the rat chooses a near-optimal route that passes through x_i and x_j.

If we didn't include structural correspondences that are approximately instantiated, then the existence of an exploitable structural correspondence would be a very demanding constraint indeed. The definition of exploitable structural correspondence puts strong restrictions on the relations that are candidates on both sides of the correspondence. In the real world it will almost never be the case that there is an exact correspondence between these relations. Requiring that such a tightly restricted correspondence should be exactly instantiated, in order for our theory to have recourse to it, would be an overly demanding constraint.

However, once we allow approximate instantiation we open up a whole class of candidate exploitable structural correspondences. Distance is a relation of significance to the rat, but co-activation maps smoothly to distance in continuum-many ways, placing different metrics on the locations represented. Which of these mappings gives the content? We answer that by looking for relations that play an unmediated role in explaining S's performance of task functions, allowing for approximate instantiation. For each exploitable structural correspondence I, we can ask how approximately or exactly it was instantiated across the range of cases that were involved in stabilizing **(p.141)** the system's task functions and producing them robustly. Suppose I maps $V(v_i,v_j)$ to $H(x_i,x_j)$. We can consider all the occasions when tokening of representational vehicles figures in the explanation of task functions and calculate how closely the actual relation between x_i and x_j matches H across those occasions (e.g. how nearly their actual spatial separation matches the distance H given under *I*).[21] The sum of those values across instantiations (possibly weighted for their significance) measures how accurately or approximately I was instantiated across those occasions.

By repeating this process for all the many candidate correspondences, we get a measure for each. Generally, instantiation being less approximate will make a correspondence a better candidate for being a UE structural correspondence. But just as the content-constituting correlation needn't be the one that maximizes accuracy (Godfrey-Smith 1989), the content-constituting correspondence needn't be the least approximate one. We are in the business of explaining robustness and stabilization, so the degree to which the UE structural correspondence is approximately instantiated during stabilization should match the extent to which episodes of behaviour did eventuate in stabilization-producing feedback. As well as metrical changes, there are also correspondences with different degrees of determinateness. There is a mapping of co-activation to precise distances (e.g. 12.4 cm apart) and another mapping to determinables like *far apart* and *quite close*. Here too we are looking for a degree of determinacy that matches the degree to which instantiation contributed positively to stabilization.[22] These considerations may not settle on a unique candidate, but only arrive at a family of equally explanatory UE structural correspondence relations (e.g. with slightly different metrics), in which case there will be an equivalent degree of indeterminacy in the content.

The degree of approximate instantiation is only a subsidiary consideration in homing in on UE structural correspondence. The primary concern remains finding a correspondence with objects and properties that figure directly in a causal explanation of robustness and stabilization: how robustly produced outcomes had consequences in the world that produced effects on the organism which stabilized this behavioural tendency and the mechanism which produced it. Approximate instantiation comes in when we are deciding between different mappings to explanatory objects and properties, for example different metrics for mapping temporal differences in neural firing onto spatial differences between locations. Suppose, hypothetically, that a mapping from rat place cell co-activation to light intensity differences was more accurately instantiated than the mapping to space when its task functions were stabilized. That mapping would be a less good candidate because light intensity differences could only provide a mediated explanation of spatial route-finding behaviour. Locations, distances, and **(p.142)** rewards at locations figure directly in a causal explanation of how rat navigation behaviour is stabilized. Light intensity could only be explanatory because it correlates with these causally relevant properties.

This way of handling the approximation inherent in occasions when a real organism performs real behaviour can, I think, also handle representational redundancy. The definition of structural correspondence we have been working with follows the mathematical notion of homomorphism. Since the mapping need not be one-to-one, two vehicles may be mapped to the same entity (e.g. v_i and v_j both to x_i).[23] But suppose two place cells map to the same location, and that one activates the other. Co-activation would then represent that they are some small distance apart, which of course cannot be the case if they both map to the same location. So, this would be a case where the relation represented under the mapping (being a small distance apart) is only approximately instantiated on the occasions that go into explaining task functions (where there is no distance between the locations mapped, since they are both the same location). Thus, representational redundancy will increase the extent to which a structural correspondence is only approximately instantiated, but mappings that contain some redundancy are not excluded from being candidates for a UE structural correspondence. Similarly, we can compare approximateness between correspondences under which the mapping of vehicles v_i to worldly entities x_j has been permuted.

(c) Evidential test for UE structural correspondence

The idea of approximate instantiation gives us another useful tool. When discussing UE information in the last chapter, I suggested a rough-and-ready evidential test (§4.2). The correlation whose strengthening and weakening is most directly tied to the likelihood of the system achieving its task functions is a good candidate to be UE information. We now have the tools to formulate a similar evidential test for UE structural correspondence. Here we look at how accurately or approximately a correspondence obtains on occasions when it is instantiated. We then apply the same idea. For a candidate structural correspondence I, we see what the effect would be if it were instantiated more accurately. Would the system be more likely to achieve its task functions? A correspondence for which the accuracy of its instantiation is more directly connected to the likelihood of achieving task functions is a better candidate for content.

> *Evidential test for UE structural correspondence*
>
> The exploitable structural correspondence defined on putative vehicles of content in a system S performing task functions F_j which is such that
>
> being less approximately instantiated most increases and being more approximately instantiated most decreases the likelihood of S achieving F_j

is a good candidate to be a UE structural correspondence.

(p.143) As we saw before, this test may be empty, indeterminate, or of little practical use. But it will often home in on content. Something like it is often at work epistemically in assigning content in real cases in cognitive neuroscience. The varitel framework allows us to see why that should be. The test also helps with some of the questions about indeterminacy we saw in the last section. Rat place cells have a less determinate and a more determinate mapping to distance (quite far away vs. 22.4 cm away). The evidential test counts against the less determinate mapping.

As before, the test is only applied to objects and properties in the world that are of significance to the organism, so it is in some ways subsidiary to constraints deriving from causal explanations of task functions. It is important to note that it does not imply that content is given by the most accurate correspondence (the one which is least approximately instantiated). It tests for how much changes in accuracy would impact on the likelihood of S producing task functions F_j and receiving stabilizing feedback. For example, prey animals frequently dart away from noises. The occasions which contribute to stabilization, that is when a predator is present, are much rarer (Godfrey-Smith 1991). However, on those occasions it is the relation with predators that has the most direct effect on whether the hapless prey achieves the task function of avoiding predators.

To see the test in action, let’s revisit the experiment performed by Constantinescu et al. (2016). They obtained evidence that subjects had learnt a two-dimensional space for a series of cartoon birds, with the dimensions defined by leg length and neck length. They found that distance N in neural activation space corresponds to similarity S_{2D} in that two-dimensional feature space. Furthermore, this correspondence explains how subjects are able to move from a starting state to a target image with the minimal amount of adjustment. Now consider a different (but closely related) candidate structural correspondence: the correspondence between neural activation distance N and the leg-length dimension taken alone S_{1D}. How accurately or approximately neural distance N mirrors leg-length would also have an impact on the likelihood of the subject achieving an efficient adjustment to reach the target image. But it would have less of an effect on achieving that outcome than the correspondence between N and S_{2D}. Consider another even weaker candidate: the correspondence between N and the overall size of the image. Instantiating that correspondence more accurately when performing the task would have a negligible effect on task performance, it might even decrease it. So, in this case the evidential test plausibly picks out the UE structural correspondence (the 2D feature space).

5.8 Conclusion

Representations are stand-ins. What better stand-in than symbols that are isomorphic to the domain you are reasoning about? It is a small step from that observation to the claim that content is based on isomorphism, homomorphism or structural correspondence. That cannot just be a matter of first order resemblance, but once we cast the **(p.144)** net more widely, it is unclear where to stop: the standard objection is that second-order resemblance, isomorphism, and other correspondence relations are too liberal to contribute any substantial restriction to a plausible theory of content. If content were ubiquitous it would lose its explanatory purchase. From our perspective—in which a theory of content is constrained by the explanatory role of representation—this liberality is a symptom of a deeper problem. The vast majority of structural correspondences which exist are not usable. And even where there is an obvious and sometimes exploitable structural correspondence, it is often not being used by the system doing the representing. On the other hand, where a system is systematically sensitive to a relation on a collection of vehicles, having that relation correspond to a relation in the world that matters to the organism—that is significant for the performance of task functions—is a very substantial achievement indeed. In this chapter we saw that instances of such an exploitable structural correspondence can take centre stage in explaining how an organism performs task functions. By this route, structural correspondence is a basis of content: it is a necessary part of a sufficient condition for content determination.

Notes:

([1]) More generally, structure-preserving. Here we focus on relation-preserving correspondence.

([2]) There are confusingly many relations in play. The exploitable relation is the correspondence, not the relations that are preserved under the correspondence.

([3]) There is parallel evidence in the human brain of similar kinds of preplay firing of sequences that correspond to trajectories through space (Horner et al. 2016, Bellmund et al. 2016); and for grid cells in entorhinal cortex, which also show prospective activity in the rat (de Almeida et al. 2012) which is coordinated with place cell activity at rest (Kropff et al. 2015).

([4]) Most models envisage a diffusion process that starts at the place cell associated with reward and proceeds outwards in parallel across connected locations (Ponulak and Hopfield 2013, Khajeh-Alijani et al. 2015). E.g. Reid and Staddon (1998, 1997) have a model in which a value signal diffuses in parallel across an array of place cells, resulting in local signals of the direction of a shortest route to a goal (discussed by Godfrey-Smith 2013). Samsonovich and Ascoli (2005) construct a connectionist model in which relations of phase precession between place cells are used to search through routes in parallel, in a ‘fan’ proceeding outwards from the current location to all nearby locations. And Corneil and Gerstner (2015) construct an attractor network where associations between place cells constrain activity directly so that the offline preplay sequence spontaneously follows the shortest route to reward in the place cell activation space.

([5]) Notice that there is no straightforward consumer for the offline activity of a single place cell. Its activity has to interact with the activity of many other place cells. The result is then used, in conjunction with other inputs about current location, to condition behaviour.

([6]) If the homomorphism is not an isomorphism, then the relation H between worldly entities needs to be reflexive, at least for entities that are mapped to by two different vehicles. If the structural correspondence maps v_i and v_j to the same x_k, then relation H has to obtain between x_k and itself. For example, relation H could be *being less than 5 cm away*.

([7]) The definition can readily be generalized to cover any collection of relations and operations, of any polyadicities, following the mathematical definition of a relational homomorphism (although the latter are usually thought to range over mathematical objects).

([8]) There is a parallel here with exploitable correlations. It is useful to have a system that can create exploitable correlations by building associations between existing correlation-carriers, e.g. a new C that is active only if A and B are. The existing correlation-carriers A and B give the system the potential to track C. Only once the new correlate is created, however, is there a new exploitable correlation. (And it is still a further step, of course, for the system to make use of that exploitable correlation.)

([9]) As before (§4.2a), being ‘unmediated’ is intended to rule out cases where I is explanatory because its targets fall under another structural correspondence I* with further objects and properties that are the ones which figure in a causal explanation of stabilization and robustness.

([10]) I won’t attempt a careful treatment here of the appropriate compositional semantics for maps, e.g. whether absence of a symbol at a location represents absence of the corresponding property instance at that location. See Blumson (2012), Camp (2007), Rescorla (2009b, 2009a).

([11]) This learnt relation plays an important role in Carey’s theory of the acquisition of number concepts (Carey 2009; see also Shea 2011c).

([12]) Another obvious case to think about is predication in a natural language sentence. Predication is a relation between representational vehicles (words) and thus is a candidate to form one end of a structural correspondence. Difficulties arise when we ask what relation in the world it corresponds to. Instantiation (of a property by an object) is the obvious candidate, but then the

Bradley regress threatens. Since we set aside linguistic representation at the outset, I won't get into those difficulties here.

[13] As in Constantinescu et al. (2016). In that case the dimension of similarity was objective, i.e. not determined by the way people tend to judge or experience the objects' similarities and differences. However, the property could equally be intersubjective, i.e. dependent on how people in general tend to experience the objects (so not fixed by similarity and difference in this subject's individual responses). If the task involves coordinating with others (e.g. in picking a colour scheme), then feedback, hence stabilization, depends on the individual's similarity judgements accurately tracking this intersubjective response-dependent dimension of similarity.

[14] This is a simplification. It would be more realistic to suppose that recognizing objective similarity and difference is a means to performing some different task function.

[15] If the neural activation space arises as a result of training, as in neural network models, then this is another case where the exploitable structural correspondence arises at the same time as it is stabilized (§5.4a).

[16] In many causal learning experiments, subjects have to learn about causal structure during reinforcement, i.e. while they are learning how to behave in reliance on the structural correspondence which is being created, e.g. Goodman et al. (2007). So, these are further cases where the exploitable structural correspondence comes into existence while it is being stabilized (cp. §5.4a).

[17] In this experiment the repeatable patterns were measured across the whole brain. The hippocampus alone is unlikely to be coding the images directly, but may be coding the position of an image in a sequence: a distributed pattern of firing specific to an object at a location can be decoded from hippocampal activity (Hsieh et al. 2014).

[18] Karen Neander's recent book makes second-order resemblance constitutive of content in some cases (e.g. for perceptual states); however, she sees second-order resemblance as a supplement to her causal-teleosemantic theory (Neander 2017, pp. 175–215), where I take structural correspondence to be an alternative basis of content. (Also, my notion of structural correspondence is not limited to relations that meet the conditions for being a similarity/distance relation.) Neander's account has the same attractive consequence that it fixes content for new representational vehicles that fall under the same relation.

[19] The repeating hexagonal pattern of grid cells is another candidate (Constantinescu et al. 2016). This structure can be used for its correspondence

to relations in the world with the same structure, when that is relevant to a new task, even though that is not why the neural structure exists.

([20]) Defined at the end of §5.4a above.

([21]) In cases considered at the end of §5.4 above, where the referent of the vehicles v_i and v_j is not already fixed (e.g. by UE information), we also need to consider how permutations of their referents would affect accuracy.

([22]) Optimality is a special case of this. One approach to representation in cognitive science is to lean heavily on optimality. Organisms are said to represent contents that make them cognitively optimal in some sense (e.g. Bayes rational). From our perspective that is a special case of this more general principle.

([23]) I.e. homomorphism allows functions that are not surjective.