Cognitive Models are Distinguished by Content, Not Format

Patrick Butlin

King's College London

Abstract

Cognitive scientists often describe the mind as constructing and using models of aspects of the environment, but it is not obvious what makes something a model, as opposed to a mere representation. The leading proposal among philosophers is that models are structural representations, and are therefore distinguished by their format. However, an alternative conception is suggested by recent work in artificial intelligence, on which models are distinguished by their content. This paper outlines the two conceptions, and argues for the content conception, against the standard philosophical view.

1. Introduction

In cognitive science, the mind is often described as working by constructing, manipulating and exploiting models of the body and the environment. This idea was introduced by Craik (1943) and has become hugely influential.[1] But what exactly is meant by 'models' in this context? What distinguishes the claim that the mind *models* entities with which it interacts, from the presumably weaker claim that the mind *represents* such entities? Recent philosophical work has linked the concept of a model to that of structural representation, often focusing on the hierarchical generative models of predictive processing (Gładziejewski 2016, Gładziejewski & Miłkowski 2017, Kiefer & Hohwy 2018, Williams & Colling 2018). This connection has also been made in older work, such as by Cummins (1989), Ryder (2004) and Ramsey (2007). So one possible view is that modelling is the use of specifically structural representations, rather than representations with some other format. In this paper my aim is to put forward an alternative view: I will argue that models should be distinguished from other representations by their content, not by their format.

More precisely, my aim is not to argue that philosophers and cognitive scientists never use the term 'model' to mean 'structural representation', or that they would always be wrong to do so. Instead, I want to draw attention to two points. First, the phenomena that philosophers and cognitive scientists call 'models' are typically structural representations, but they *also* have content of a certain kind, which distinguishes them from other representations. The generative models of predictive processing, for example, have both of these features. This raises questions about the relationship between the features – for example, whether the use of a structural format is particularly apt for representations with content of this kind – and also about the different ways in which they facilitate cognition, which may be obscured by talk of models which is insensitive to the difference between the two features. Second, there are examples of representations used in algorithms of quite central importance in cognitive science, which are naturally and universally referred to as models, but which do not have a structural format. So having a structural format does not seem to be necessary for a representation to be a model.

I also certainly do not wish to deny that there is an important and interesting distinction between structural representations and those with other formats. Indeed, my argument relies on this distinction. My substantive claim is that there is a *further* important and interesting way to

[1] In addition to the examples I discuss below, the idea that the mind uses models has been prominent in theories of reasoning (Johnson-Laird 1983, 2006) and motor control (Wolpert et al. 1995, Grush 2004). Webb & Graziano (2015) invokes models in their theory of consciousness, and Danks (2014) argues that diverse cognitive processes are united in their reliance on graphical models. But these are merely a few prominent examples; cognitive scientists refer to models very frequently.

classify representations used in cognitive systems, which does not correspond exactly to the structural/non-structural distinction, but which captures part of what scientists intend when they speak of 'models', in a range of significant cases.

One complication is that talk of models contributes to modern computational cognitive science in two ways (Eliasmith 2007). As well as claiming that the mind models the body and the environment, cognitive scientists construct models of the systems they study. They engage in 'model-based science' (Godfrey-Smith 2006a) by, for instance, developing computer learning algorithms and testing their performance on cognitive tasks. There is an extensive literature on the role and nature of the models used by scientists of all disciplines (e.g. Giere 1988, Magnani & Nersessian 2002, Weisberg 2013), and this includes debate on the means by which scientists' models represent their targets, in which structural accounts have taken a prominent place (Suárez 2003). However, I must emphasise that my topic is not scientists' models. My topic is the mental models which humans and animals are said to use for a wide variety of cognitive tasks.[2]

This paper has three main parts. In the next section, I define structural representation and give more detail on the claim that modelling is the use of representations of this kind. In section 3 I present the alternative, content-focused conception of the use of models in the accomplishment of cognitive tasks. Then in section 4 I argue against the structural representation view, using the example of a model-based reinforcement learning algorithm which may be implemented in simple computer programs.[3]

2. The Format Conception

In this and the following section I describe two possible conceptions of a cognitive model, which I call the 'format' and 'content' conceptions. This section presents the format conception, in three stages: first I clarify the notion of representational format; then I explain what is meant by 'structural representation'; and finally, I state the format conception and summarise the reasons why philosophers have adopted it.

2.1 Representational Format

In claiming that models are distinguished by their content rather than their format, I am claiming that one dimension of variation among representations is implicated in this distinction,

[2] Scientists' models are important here in one way, which is that I will appeal to computer programs used in reinforcement learning research – which might be thought of as scientists' models – in arguing against the conception of mental models as structural representations. I discuss this complication further in section 4.

[3] Note that the practice of calling these two kinds of reinforcement learning 'model-based' and 'model-free' is entirely standard in the discipline – my use of these labels is not prejudicial.

rather than another. However, there may be some uncertainty about what I mean by the 'format' dimension, and clearing this up will also help to make clear how I understand the notion of structural representation.

In my way of thinking about things, representations vary along at least three dimensions. The first is content, which is sufficiently familiar that I won't say anything more to explicate it. The second is format, and the third is what I shall call 'basis'. A representation's format is the way in which its vehicle properties are used to perform its representational function. Maps, photographs, sentences and musical scores are said to employ different formats, and in each of these cases vehicle properties are used for representation in different ways. The basis of a representation is the set of properties of the system within which it is embedded that make it the case that it has a representational function, and that its vehicle properties are used to represent in a particular way. The basis of a representation might be, for example, a convention that governs its use, or the biological function of a system that employs it.

To consider an example in slightly more detail, one philosophical debate which is explicitly about representational format concerns whether there is a proprietary format for perceptual representation (Quilty-Dunn forthcoming). Some contributors to this debate argue that perception has an *iconic* format, meaning that perceptual representations do not admit of canonical decompositions (Carey 2009, Burge 2010, Block 2014). That is, any part of a representation in iconic format is equally representational, as in a typical photograph. This contrasts with non-iconic representations such as sentences, which have parts that do represent (such as words), and others that don't (such as part-words). Quilty-Dunn (2016) argues that at least some perceptual representations are not iconic, by arguing that objects are sometimes represented by discrete representational constituents, which behave more like words than like the parts of a photograph.

This debate fits my definition of format, because it concerns different ways in which vehicle properties can be used to perform representational functions. If the iconic theory is correct, then in perception representational vehicles, instances of brain activity, are used in such a way that each of their parts plays equally representational roles. Perhaps any part of an instance of brain activity constituting a moment's visual perception represents the distribution of colours over a region of the visual field. On the other hand, if Quilty-Dunn is right, then there are certain parts of the representational vehicles used in perception that represent objects, and are such that their proper parts do not represent anything at all.

Representational basis should be distinguished from format, because representations can be similar in format and different in basis, or vice versa. Consider a weather map that represents

land-masses by scale line drawings, and forecast temperatures by coloured regions. Such a map is an iconic representation, so if the iconic theory is right it is similar in format to perceptual representations. These differ in basis, however, because what makes the weather map a representation, and grounds its format, is the conventions governing its use. Perceptual representations, meanwhile, have their basis in either biological functions established by natural selection, or something like Shea's (2018) 'task functions'. Conversely, if Quilty-Dunn (forthcoming) is right that perception uses a plurality of representational formats, then it is possible for representations with the same basis to differ in format.

Two further points about format are important for my purposes. First, the concept of *exploitable relations* (Godfrey-Smith 2006b, Shea 2018) is sometimes invoked in connection with structural representation, and this is helpful in one way, and unhelpful in another. Shea (2018) argues that subpersonal mental representations can bear either of two exploitable relations to the phenomena they represent: they can either carry information about those phenomena, or correspond to them structurally. Representations that differ in which of these relations are actually exploited by cognitive systems consequently differ in format, because their vehicle properties are used for representational purposes in different ways. So this point is very helpful in distinguishing structural representations from others. But the idea can also be confusing, because there is room for argument about whether all representations work by exploiting exploitable relations between vehicles and the phenomena represented. In particular, this is less clear in the case of representations with a conventional basis. The upshot is that if a representation exploits an exploitable relation, such as structural correspondence to its target, this has consequences for its format; but it is clearer that all representations have a format than that they all take advantage of exploitable relations.[4]

Second, there are many ways to classify representational formats. Classifications can be more or less fine-grained, and different schemes may give overlapping classifications, or perhaps even orthogonal ones. I take structural representation to be a broad class of formats, so some intuitively distinct formats (such as maps and photographs) may be brought together under this classification.

[4] O'Brien and Opie (2004), following Von Eckardt (1993), suggest that representation can be 'grounded' by either resemblance, causation or convention. But this classification, which combines format and basis, is problematic because mere resemblance or causation is not sufficient for representation without a basis such as convention or biological function, and because some representations both exploit resemblance, and have a conventional basis. Hieroglyphs are one example among many.

2.2 Structural Representation

Structural representations (henceforth SRs) are a species of representation defined by similarity in their formats. More specifically, SRs have the following two defining properties:

i. Relations over parts or states of the representational vehicle are used to represent relations over parts or states of the represented phenomenon.
ii. The relations over the vehicle are used in this way because there is a putative structural correspondence between them and the relations over the represented phenomenon.

This definition draws on discussions of structural representation by Swoyer (1991), O'Brien and Opie (2004), Ramsey (2007), Shagrir (2012) and Shea (2014, 2018). Shea defines structural representation only in terms of the first of the two properties, but makes clear that he also takes it to require structural correspondence. O'Brien and Opie do not use the expression 'structural representation', but theorise about a form of representation that relies on structural correspondence, defined in terms of structure-preserving mappings, which is how I define this concept below. Swoyer and Shagrir both define structural representation in terms of structural correspondence, and Ramsey describes SRs as relying on structural similarity, correspondence or isomorphism. The point that the structural correspondence between the vehicle and the representated phenomenon is merely putative is rarely mentioned, but necessary because of the possibility of inaccurate SRs, where the correspondence fails.

Following Shea and others, I define structural correspondence in terms of homomorphism, understood in the following way.

> Let R_A be a relation over parts or states of A, the representational vehicle, and let R_B be a relation over parts or states of B, the thing represented. Then f, a function from parts or states of A to parts or states of B, is a homomorphism with respect to these relations if and only if: for any pair a_1 and a_2, $a_1 R_A a_2 \leftrightarrow f(a_1) R_B f(a_2)$.

For a vehicle to bear a structural correspondence to a represented phenomenon is just for a homomorphism to exist from the former to the latter. Homomorphisms are abundant, so to say that structural representation requires the existence of a homomorphism from the vehicle to the represented phenomenon is a very weak constraint. When combined with the first condition on structural representation, however, the constraint becomes significantly stronger.

Maps are often given as examples of SRs (Shea 2014, Gładziejewski 2016), and indeed they satisfy the definition. For example, consider the famous map of the London Underground network. This map has certain privileged parts, the marks symbolising each station, which correspond in the obvious way to the stations themselves. A function taking each of these marks to the named station is a homomorphism with respect to some relations of particular salience on the map, and some corresponding ones of practical importance in the world. It maps the relation *being connected by a line of a single colour* over the marks for the stations, to the relation *being connected by a single London Underground line* over the stations themselves; and it also preserves the orders in which stations are thus connected.

A crucial point for understanding the notion of structural representation is that for an entity to be an SR, a set of relations over its parts or states that structurally correspond to some relations over the parts or states of a further entity *must be used to represent those relations*. Many cases of structural correspondence are therefore not cases of representation at all. For example, there may be a very natural homomorphism between a set of 100 bricks arranged in offset rows in a wall, and 100 football fans sitting in similarly offset rows in a stadium, with spatial relations between the bricks corresponding to similar spatial relations between the fans, but the theory of structural representation does not entail that the bricks represent the fans.

What's more, not all representations (or collections of representations) that are homomorphic to the things they represent are SRs, because in many cases homomorphisms exist even though the relevant relations over the represented domain are not represented. Shea (2014, s. 3) illustrates this point with the case of the honeybee's waggle dance, in which the angle of the dance corresponds to the direction of the source of nectar, and the number of waggles corresponds to its distance. There is a homomorphism between dances and locations, and this homomorphism determines the representational content of any given dance. But this may not be a case of SR, because for that to be the case, relations between dances would have to condition the behaviour of observer bees in ways that would indicate that they represent relations between the locations of nectar sources. For instance, bees might watch two dances, and use the relation between them to fly first to one nectar source, then directly to another, without going back to the hive. But if bees don't do such things, merely using individual dances as guides to the location of single nectar sources, the relations between dances are not used in the way required for structural representation.

For a representation to be an SR, then, the details of the way in which it is used are crucial. The system that uses it must be capable of behaving in ways that are sensitive to relations over the represented domain, in virtue of the correspondence between these and certain relations over

the representational vehicle. This sensitivity may, however, be subtle or indirect. If an organism uses some internal process as a dynamic model of the environment which predicts incoming sensory stimulation, this model may contribute directly only to perception. But by facilitating faster or more reliable responses to environmental contingencies, the model could affect the organism's behaviour; and these effects on behaviour would constitute sensitivity to relations between features of the environment, such as the tendency for one kind of event to be followed by another.

The final aspect of structural representation which I want to discuss concerns look-up tables. Representations of this form play an important role in the argument to come, and there are some subtleties in how the definition just given applies to them. It will be helpful to have an example in mind, so consider a table of this form, into which many of the facts represented by the London Underground map could be transcribed:

Departure Station	**Destination Station**	**Line**	**Stops**	**Direction**
Acton Town	Aldgate East	District	22	East
Acton Town	Alperton	Piccadilly	4	West
…	…	…	…	…
Temple	Westminster	District	2	West
…	…	…	…	…

The table is in alphabetical order, and lists each pair of stations connected by a single line, which line connects them, and how many stops are required, in which direction. Intuitively this is an example of a way to re-represent much of the information in an SR in a non-structural format. But in fact matters are slightly more complicated; I will note three points about how the definition of an SR applies to this case, and to other look-up tables.

First, the table certainly employs what might be called *non-structural elements*. It identifies stations and lines by their names, and this way of representing does not make use of structural correspondence. Furthermore, it even represents some worldly relations other than by relations over the vehicle: both the number of stops between stations, and the direction of travel between them, are represented by symbols rather than relations. The use of non-structural elements is common in artifactual SRs such as maps, especially to identify objects, and this should not prevent us from recognising the crucial role that structural correspondence may still play. Unless we find a good reason to do otherwise, we should count all representations that satisfy the two conditions above as SRs, even if they also employ non-structural elements.

Second, this look-up table does have one of the two defining features of SRs, because the vehicle relation *being on the same row of the table* is used to represent the worldly relation *being connected by a single London Underground line*. Something similar will also be true of many other look-up tables. However, there is no structural correspondence between these two relations, because the function from station names on the table to stations themselves is not one-to-one, but only one token of each name on the table stands in the relevant vehicle relation to each other name. Going more slowly, let us apply the definition of a homomorphism given above to the present case in the following way:

Let: $a_1, \ldots, a_n$ be token names on the table
$b_1, \ldots, b_m$ be stations on the London Underground network
R_A be the relation *being on the same row of the table*
R_B be the relation *being on the same London Underground line*.

Then the function mapping station names on the table to the stations that they name is not a homomorphism with respect to these two relations, because it is not true that for any pair a_1 and a_2, $a_1 R_A a_2 \leftrightarrow f(a_1) R_B f(a_2)$. Consider the first token of the name 'Acton Town', and the token of the name 'Westminster', on the fragment above. If these are a_1 and a_2 respectively, then $f(a_1) R_B f(a_2)$ is true, because Acton Town and Westminster are both on the District Line. But $a_1 R_A a_2$ is not true, because these two tokens do not appear on the same line of the table. So the look-up table as a whole is not an SR. Its structure does not correspond to the structure of the system it represents, because relations over that system are represented piecemeal. Where in the system represented each object stands in many relations to other objects, in the representational vehicle there are no parts that stand in the corresponding set of relations.

This pattern is likely to be common where information is stored in look-up tables. Part of what defines a look-up table is that the rows and columns have representational significance, but there are few structures of relations that can be represented in a look-up table other than in the piecemeal way just described.

Third, despite this, each row of the table we are considering is an SR, according to the definition. The function that maps each station name on the first row of the table to the corresponding stations is a homomorphism with respect to the two relations *being on the same row* and *being on the same line*. This shows that a composite representation made up of parts which are SRs will not always be an SR itself. It also shows that not all SRs make any very interesting use of structural correspondence. If each row on the table we considering were replaced by a sentence

of English, the same information would be represented with very little change in efficiency or accessibility. Yet English sentences are not SRs; for example, in the sentence 'Acton Town is on the same line as Aldgate East' the relation of being on the same line is represented by a part of the vehicle, not a relation over parts.[5] There is a stark contrast between the rows of our table and SRs which do make significant use of structural correspondence, such as the London Underground map itself. It is thanks to its use of structural correspondence that this map stores and presents information in a highly efficient and accessible way.

2.3 Arguments for the Format Conception

Now that we have the notion of structural representation in hand, we are ready for a statement of the format conception. The format conception is the claim that cognitive models are structural representations.

> **Format Conception:** A cognitive model is a structural representation used in performing a cognitive task.

Since we know what SRs are, it remains only to briefly consider arguments for the format conception.

The format conception of cognitive models is motivated by at least three lines of thought. First, many artifacts which we refer to as 'models' are used as SRs. For instance, Ryder (2004) mentions dynamic models of the solar system, which can be used to answer questions about possible spatial relations between the planets. Swoyer (1991) mentions a model aeroplane used for wind-tunnel testing. This point certainly makes it natural to use the term 'model' to describe at least some SRs in cognitive systems. Second, it is sometimes argued that causal and informational theories of mental representation (also called 'indicator' or 'detector' theories) suffer from insuperable difficulties, which can be avoided if we explain representation in terms of exploitable structural correspondence (Ramsey 2007, Williams & Colling 2018). So the

[5] Relations over the parts of English sentences do contribute to determining content; 'John loves Mary' means something different from 'Mary loves John'. So there are composite representations which are not SRs, but which are such that relations between their parts contribute to determining content. Kiefer and Hohwy (2018) also make this point, in comparing theories of structural representation to functional role theories of mental content. In the case of sentences, it's noteworthy that relations of concatenation do not generally represent worldly relations. For example, in the sentence 'Alice runs' the concatenation between 'Alice' and 'runs' does not represent a relation between Alice and the property of running. To say that it did would launch us on a version of Bradley's regress (Bradley 1893): if it is necessary to represent relations between objects and properties, it must also be necessary to represent the relations between objects and those relations, and so on.

thought is that by showing that cognitive models are SRs, it is possible to defend their status as representations. And third, it has been argued that cognitive models of some specific kinds, especially the generative models implicated in the predictive processing theory of cognition, are in fact SRs (Gładziejewski 2016, Kiefer & Hohwy 2018; for the predictive processing theory see Clark 2013, 2016, Hohwy 2013). Gładziejewski's argument for this claim works by comparing these models to cartographic maps, which are taken to be prototypical non-mental SRs.

These arguments are persuasive, but they are not conclusive. The first and third arguments strongly suggest that models and structural representation are connected, but they do not show that an alternative conception of cognitive models could not be more illuminating; and the second argument relies on a highly contentious claim about the prospects of indicator theories. So I now turn to the content conception of cognitive models.

3. The Content Conception

The content conception builds on the work of Lake and colleagues (2017), who seem to rely on differences in representational content, rather than format, in drawing a distinction between what they call 'pattern recognition' and 'model-building' algorithms for cognitive tasks. Lake et al. argue that in order to construct human-like artifical intelligence we must develop algorithms of the latter kind (see also Tenenbaum et al. 2011, Garnelo et al. 2016). They suggest that only model-building cognisers can perform tasks with understanding, and apply their existing knowledge of the world effectively in learning to perform new tasks or adapting to changing environments. If they are correct, then the content conception of a model may contribute to distinguishing between truly intelligent artificial systems and programs which perform impressively on cognitive tasks only through brute force or clever tricks. In essence, Lake et al.'s distinction is between one group of algorithms which learn to proceed in a single step from input to output (the pattern recognition type), and another which constructs and employs structures which represent underlying features of the domain with which the system interacts (the model builders). This is intended to capture common features of two distinctions: between model-free and model-based reinforcement learning, and between discriminative and generative classifiers.

I will discuss these two examples in turn, and argue that in both cases, the model-building algorithms employ representations with content which is apt for use in explaining the inputs they receive and for justifying outputs. In the two examples, the representations in question have more in common than just this: they also represent relations which make it possible to calculate the probabilities of states of the environment, and of forms of input given these states. But for

reasons I will explain further below, I think that specifying the content conception in terms of aptness for explanation and justification is more likely to give a fully general characterisation than any more direct account of the content involved. For the avoidance of doubt, my claim is not that the use of a cognitive model requires the *capacity to give* explanations or justifications, but only that it requires the use of representations with *content that could contribute* to explanations or justifications. So on this account what makes a representation a model will indeed be its content, not the cognitive functions for which the system in question can use it.

In the rest of this section, I first discuss the two examples, then present the content conception, and draw out an important implication.

3.1 Two Examples

In model-based reinforcement learning (RL), agents learn to select rewarding actions by coming to represent two kinds of information about their environment. They learn about the values of possible outcomes, and about relations between actions and outcomes. To choose the most rewarding action in a given situation, agents using this system may use these representations to generate a decision tree, showing the outcomes accessible from their present situation by a single action, the outcomes accessible from each of those, and so on; and then to evaluate each possible course of action on the basis of the value of the sequence of outcomes it will bring about. In contrast, in model-free RL agents learn only about the levels of reward brought about by actions in given situations, without learning which outcomes result from these actions. In effect, they learn the values of actions, rather than of outcomes, and they do not learn about relations between actions and outcomes. Model-free algorithms learn directly what output should be produced (the action) in response to each input to the system (the situation); model-based algorithms learn facts about the environment which allow them to reason about which action to perform.

Model-free RL algorithms such as temporal difference learning can learn the long-run values of actions (Sutton & Barto forthcoming), thus avoiding the computational costs associated with generating and evaluating decision trees. But they are inflexible compared to model-based algorithms in the following sense: model-based algorithms can quickly adapt to small changes in outcome values or action-outcome contingencies, by updating the corresponding representations; but model-free algorithms must re-learn policies from scratch to accommodate such changes. As Lake and colleagues explain, a human video game player could immediately perform competently in a new version of a game with a different objective or pattern of rewards and punishments, but Mnih et al.'s (2015) deep neural network which achieved human-like

performance on a range of Atari games requires a great deal of retraining to cope with such changes (Rusu et al. 2016), because their system relies solely on model-free RL.

This difference in flexibility and the ability to apply knowledge in new conditions is illustrated by the standard experimental methods for distinguishing model-based from model-free action selection. In one method, *outcome devaluation* (Balleine & Dickinson 1998), animals learn to perform an action, such as pressing a lever, for a reward, which is typically an unfamiliar food. The reward is then devalued, away from the setting in which the action has been learnt. For example, the food may be paired with an injection of a substance that causes gastric illness. The animals are then tested to see whether they resume performing the action in the original setting, without the reward being delivered. Continuing to perform the action is considered to be evidence of model-free RL, because this indicates that the action itself is represented as rewarding. Reduced performance is evidence of model-based RL, because it indicates knowledge that the action leads to the now-devalued reward.

A second method, the *two-step task* (Gläscher et al. 2010, Lee et al. 2014) involves participants making two choices in sequence. Each choice leads to a subsequent state with a certain probability, and reward depends only on which state is reached after the second step. There are different ways to use tasks of this form to test for model-based control, but a simple one is to allow participants to explore the state space in the absence of reward first, then inform them about which states are rewarding, and introduce the rewards. Participants who perform above chance when rewards are first available must be using model-based RL, because no actions have previously been rewarded, so a model-free system would have learnt nothing (Gläscher et al. 2010).

There are two abstract features of model-based RL in virtue of which it may be said to use models, while model-free RL does not. First, model-based RL represents relations which model-free RL does not, and which crucially go beyond the relations between inputs and suitable outputs that are most directly relevant to the task of action selection. Model-based RL represents how the environment will change over time, contingent on possible actions, and makes use of these representations to select actions. Model-free systems do represent current states of affairs, and the range of actions which are presently available, but they do not represent possible transitions between states of the environment, which are relations between such states. Model-free systems represent only the expected levels of reward of actions in each possible state; which is to say that they represent only the suitability of available outputs for each possible input.

Second, these relations allow the formation of expectations about environmental contingencies, and are apt to faciliate hypothetical reasoning, and to ground explanations of

inputs and justifications of outputs. Both model-based and model-free RL algorithms are good for returning representations of actions which are likely to be rewarding as outputs when fed representations of states of affairs as inputs. But model-based RL also has the resources to perform further related tasks. It employs representations which have the potential to be used to predict what will follow from either the current state of affairs, or some hypothetical alternative, given each of a range of possible actions. This information can be used to give non-trivial justifications of selected actions – model-free RL says only that the chosen action is the best available, whereas model-based RL says what is good about it.

Turning now to the distinction between discriminative and generative classifiers (Lake et al. 2017, Jebara 2004), a typical task in machine learning is to classify inputs, such as handwritten characters. In this context inputs are sometimes called 'data', and outputs are 'labels'. Discriminative classifiers are those that work by learning and applying a representation of the probability distribution of labels given data. These representations allow them to perform classification tasks directly; given a particular handwritten numerical character, the system will read off from the distribution the probabilities that the correct label for this shape is 0, 1, 2, … or 9, and can simply pick the label with the highest probability. Generative classifiers learn the distribution of data given labels, and the prior probabilities of labels, which enables them to perform the task thanks to Bayes' rule. Compared to discriminative algorithms, they work the 'other way around'; they attempt to match the data to representations of likely shapes corresponding to each character. In a sense, a generative algorithm for classifying handwritten characters relies on knowledge about what each character is like, rather than knowledge about which shapes constitute which characters.

As I mentioned above, generative models are central to the predictive processing theory of cognition (Hohwy 2013, Clark 2013, 2016). As it applies to perception, this theory claims that we perceive the world through an ongoing process of hypothesis-formation and –testing. The way that I represent the world as being at time t causes my perceptual system to represent a probability distribution over ways the world might be at time t+1, from which a hypothesis is formed. This hypothesis predicts that I will undergo certain sensory stimulation at t+1, and this prediction can be used to test and revise the hypothesis, until one is found which minimises prediction error. In order to perform these steps, the perceptual system must employ a generative model. It requires representations of the probabilities of forms of sensory stimulation (the input to the system) given hypotheses about how things are (the output), and also of the causal and/or informational relations between states of the world at successive times or in neighbouring locations.

Generative classification algorithms are therefore distinguished from discriminative algorithms by the same two abstract features that distinguish model-based from model-free RL, although it is the second that defines the generative-discriminative distinction. First, in some important cases generative algorithms represent relations which are not represented by discriminative algorithms. As just described, a prediction error minimisation algorithm for perception will use a representation of relations between successive states to generate hypotheses concerning incoming stimuli; without this knowledge far too many hypotheses would have to be tested for the system to perform the task effectively. In terms of data and labels, this representation of relations provides the prior probabilities of labels.

Second, unlike discriminative algorithms, which proceed directly from input to output, generative models represent features of the task domain that explain the data. These features include both relations between hidden variables, and the typical sensible qualities of the kinds which the system aims to identify. If human perceptual systems use generative models, these represent relations that have the potential to explain why we receive the sensory stimulation we do; for example, I may represent that a goal has just been scored, that goal-scoring tends to cause cheering, and that cheering sounds a certain way, and these facts will together explain the sounds I am currently hearing. The same aspects of the representational content of generative models allow the formation of expectations about stimuli, and are apt to make hypothetical reasoning possible.

3.2 Formulating the Content Conception

Model-based RL algorithms and generative classifiers have in common that, unlike their respective alternatives, they represent features of task domains which are capable of explaining their inputs or justifying their outputs. Model-based RL algorithms represent causal relationships between states, actions, and subsequent states, and these have the potential to contribute both to explaining why agents find themselves in the states they do, and justifying their chosen actions. Generative classifiers represent the probabilities of data given labels, and these representations can contribute to explaining why the data they receive take the forms they do. Generative models in predictive processing, moreover, represent causal or informational relations between successive states, which can contribute further to explaining sensory stimulation.

These features distinguish model-building from pattern recognition. So the content conception of a cognitive model is as follows:

Content conception: A cognitive model is a representation used in performing a cognitive task, which represents features of a task domain which are apt to explain inputs to the cognitive system and/or justify its outputs.

As Lake et al. put it, 'cognition is about using… models to understand the world, to explain what we see, to imagine what could have happened that didn't, or what could be true but isn't, and then planning actions to make it so' (p. 2).

A potential objection to this account is that it offers quite an indirect characterisation of the kind of content that distinguishes cognitive models. Instead of saying which features of the task domain models represent, it says only that they represent features which are apt for further tasks. Before I give my response to this objection, it will be helpful to note an important implication of the account.

The implication is that cognitive models do not have their status as such outright, but only in relation to particular tasks for which they are used; many representations are models relative to some task, but not to others. One example of this comes from the theory of forward models used in motor control (Wolpert et al. 1995, Grush 2004). Grush describes sub-systems of the motor control system, which he calls 'emulators', which take efference copies of motor commands as input, and produce representations of likely sensory feedback as outputs (I will use Grush's term to refer to these sub-systems, and the term 'forward model' for the representations they employ). He explains that one function of emulators is to allow motor commands to be corrected more rapidly than if real sensory feedback was required. He further explains that emulators might use either associatively-learnt look-up tables representing motor command-sensory feedback pairs, or more sophisticated 'articulated models', with parts that correspond to at least some of the parts of the musculoskeletal system itself, and generate representations of likely feedback by simulating the interaction of these parts. We can focus on the unarticulated case, and consider whether a forward model that consists of command-feedback pairs is a model at all, by the standards of the content conception.

If we take the task to be motor control, the answer is 'yes'. The inputs to the motor control system specify goal behaviours, and the task of the system is to produce sequences of motor commands which will generate behaviours matching these specifications. Relative to this task, the unarticulated forward model is a model, because the production of a particular motor command may be justified by the fact that the previous command is likely to cause a certain form of sensory feedback, in combination with facts about the current goal behaviour. If an agent is trying to post a card through a slot, and the forward model entails that the card will

move towards a position a little to the left of the slot, this justifies a new motor command to move the card to the right. However, relative to the *emulator's* task, the unarticulated forward model is, despite its name, not a model at all. The emulator's task is to take motor commands as input and produce representations of likely sensory feedback as output, and the representation in question links these inputs and outputs in a single step. Relative to this task, the forward model is just like the action value representations in model-free RL, or the representations of the distribution of labels given data in discriminative classifiers. In contrast, the articulated forward model that Grush suggests would count as a model relative to both tasks.

I now return to the objection that the content conception characterises models indirectly. The points just raised about task-relativity illustrate the difficulty of giving an alternative, more direct account of the kind of content that characterises models. To start with, consider the proposal that what characterises models is that they represent probabilistic relations between world states; this is an obvious common feature of the models in both model-based RL and predictive processing. This characterisation is problematic, because as in the case of motor control, the systems for which representations of this form are models may have sub-systems relative to which they are not models. For example, model-based RL systems include sub-systems with the function of taking representations of state-action pairs as inputs, and providing representations of likely outcomes as outputs. Relative to this function, some representations of relations between world states may not be models. Note that this is an implication not just of the content conception as stated, but also of the first-pass account of pattern recognition/model-free algorithms, which described them as those that link inputs and outputs in a single step.

A different possible alternative would be to characterise models as representations that link features other than the inputs and outputs of the systems concerned. This account has difficulty with model-free RL, however, because algorithms of this type do typically represent the levels of reward associated with each state-action pair, as opposed to merely maintaining representations of which action should be performed in each state. The content conception as stated does a better job of accommodating this example, because actions cannot be justified by saying that they are rewarding to a particular degree, or more rewarding than alternatives. In this context, to say that a chosen action was rewarding gives no new information to justify the choice. So I suggest that the content conception as stated offers the most promising way to capture Lake et al.'s insight.

4. An Argument Against the Format Conception

In this section I argue against the format conception of cognitive models by reference to Dyna, a simple model-based RL algorithm for ordinary computers (Sutton 1991). I claim that Dyna uses a model, but does not use structural representation. So Dyna constitutes a counterexample to the format conception: it shows that there are models which are not SRs.

Before I get into the argument, a comment on methodology. Earlier I distinguished cognitive models, which are my topic in this essay, from scientists' models of the phenomena they study. A possible concern about my argument is that Dyna and similar programs are scientists' models of cognitive systems, as opposed to cognitive systems themselves. Perhaps a mere scientists' model of model-based RL need not use a model itself. This objection is ill-founded, however, because there is no reason to deny that Dyna actually undergoes reinforcement learning. My method is therefore not unlike some scientists' use of models; I use a particularly simple example of the phenomenon I am interested in, in order to see what is essential to it.

Dyna is an algorithm for learning to exploit sources of reward in relatively simple virtual environments. These environments generate *finite Markov decision problems* (MDPs), which the algorithms must solve (Sutton & Barto forthcoming). In an MDP time proceeds in discrete steps, and at each time-step the agent finds itself in an identifiable state, and must select an action. At the next time-step, perhaps partly as a consequence of its action, the agent will find itself in a further state, and will receive a numerical reward. Then it must choose a further action. Transitions between states may be probabilistic, but the probabilities depend only on the previous state and the action selected. In finite MDPs, there are only finitely many possible states, actions and levels of reward.

In small finite environments, reinforcement learning researchers have used look-up tables to represent what is learnt in both model-free and model-based algorithms. This point is made clear in works such as Kuvayev & Sutton (1996) and Boyan & Moore (1995), which discuss the problem of extending algorithms for finite MDPs to ones in which there is continuous variability in states and actions. This is a crucial point, because as I have argued, look-up tables are not usually SRs.

In particular, the Dyna architecture maintains three look-up tables. One represents the long-run values associated with states, which depend on the values of likely subsequent states. A second represents the next states and rewards that follow from each state-action pair; this is referred to as the 'world model' (Sutton 1991). And the third represents a *policy* – a set of rules concerning what to do in each state, which is used to select actions. The algorithm causes the world model to be updated from the agent's experience of the environment, and the policy and

record of long-run state values are updated both by 'real' experience, and 'simulated' experiences generated by the world model. The model is therefore used for planning, understood as updating the policy in advance of action.

To see that the world model in Dyna is not an SR, we can apply the argument used in section 2.2 to show that look-up tables do not usually bear a structural correspondence to the phenomena they represent. The key point is that because each possible state will be represented by multiple entries in the world model, there is no homomorphism between this representation and the causal structure of the virtual environment.

For example, consider the fragment of a possible world model shown below. In this table, the presence of a given state-symbol in the right hand column represents that the corresponding state is caused by the performance of action A_1 in the state represented by the symbols on the same row in the left-hand column. It may be suggested, then, that the relation over parts of the table: *being to the right on the same row* is used to represent the relation: *causation via A_1 in the virtual environment*.

Initial State	**Resulting State Under A_1**
S_1	S_2
S_2	S_3
S_3	S_4
S_4	S_2

If we define a function that takes symbols in the table to the corresponding states of the virtual environment, however, we can see that this will not be a homomorphism with respect to these two relations. For the function to be a homomorphism, it would have to be the case that a given state symbol is to the right on the same row as another if and only if their corresponding states are linked by causation via A_1 in the virtual environment. This is not the case, because the 'S_1' on the first row and the 'S_2' on the last row are such that their corresponding states *are* linked by causation via A_1, but they themselves *are not* linked by the relation *being to the right on the same row*. Similar points will hold for world models for almost every possible environment.

As we also saw in section 2.2, if each row of a look-up table is considered to be a separate representation, these do satisfy the definition of an SR. But this hardly grounds on which to defend the format conception, because the point remains that the look-up table – the means by which the causal structure of the environment is represented – does not exploit any

correspondence with this structure. Its parts correspond in structure to parts of the environment, but only when such small parts are considered that this is near-trivial.

The upshot of these considerations is that Dyna does not use structural representation. With this point in hand, all we need to show that the format conception of cognitive models is mistaken is that Dyna uses such a model. I take the account of the content conception in section 3 to support this claim, because it shows how the distinction between model-based and model-free RL can be assimilated to that between generative and discriminative classifiers; this indicates that the term 'model' is not used idiosyncratically by RL researchers. So I conclude that having a structural format is not necessary for an representation to be a cognitive model.

5. Conclusion

Philosophers have often connected the concept of a cognitive model to that of structural representation, suggesting that what it is for a representation used in a cognitive process to be a model is for it to be an SR. I have presented an alternative conception, according to which models are distinguished by their representational content, not their format. Following Lake et al., this conception emphasises connections between model-building, expectation, explanation and justification. I have also argued that having a structural format is not necessary for a representation to be a model.

However, this does not mean that there are no important connections between the phenomena of structural representation and the use of models of task domains by cognitive systems. It could be that the use of structural representation is sufficient for modelling – although one reason to doubt this is that discriminative classifiers may use iconic (and hence structural) formats in representing their inputs. This point deserves further investigation. Even if SR is neither necessary nor sufficient for modelling, it could still be the case that there is a great deal of overlap between these categories, for deep reasons, perhaps to do with surrogative reasoning (Swoyer 1991). Nonetheless, theorists should pay more attention to content, and perhaps less to format, in seeking to understand cognitive models.

Bibliography

Balleine, B. & A. Dickinson. 1998. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37 (4-5): 407-419.

Boyan, J. & A. Moore. 1995. Generalization in reinforcement learning: Safely approximating the value function. In *Advances in Neural Information Processing Systems 7*, pp. 369-376.

Block, N. 2014. Seeing-as in the light of vision science. *Philosophy and Phenomenological Research* 89 (3): 560-572.

Bradley, F. H. 1893. *Appearance and Reality*. London: George Allen & Unwin.

Burge, T. 2010. *The Origins of Objectivity*. Oxford: Oxford University Press.

Carey, S. 2009. *The Origin of Concepts*. Oxford: Oxford University Press.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences* 36 (3): 181-204.

Clark, A. 2016. *Surfing Uncertainty: Prediction, Action and the Embodied Mind*. New York: Oxford University Press.

Craik, K. 1943. *The Nature of Explanation*. Cambridge, UK: Cambridge University Press.

Cummins, R. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Danks, D. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: MIT Press.

Eliasmith, C. 2007. How to build a brain: From function to implementation. *Synthese* 153 (3): 373-388.

Garnelo, M., K. Arulkumaran & M. Shanahan. 2016. Towards deep symbolic reinforcement learning. arXiv preprint.

Giere, R. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gläscher, J., N. Daw, P. Dayan & J. O'Doherty. 2010. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66: 585-595.

Gładziejewski, P. 2016. Predictive coding and representationalism. *Synthese* 193 (2): 559-582.

Gładziejewski, P. & M. Miłkowski. 2017. Structural representations: Causally relevant and different from detectors. *Biology & Philosophy* 32 (3): 337-355.

Godfrey-Smith, P. 2006a. The strategy of model-based science. *Biology & Philosophy* 21: 725-740.

Godfrey-Smith, P. 2006b. Mental representation, naturalism and teleosemantics. In MacDonald & Papineau (eds.), *Teleosemantics*. Oxford: Oxford University Press.

Grush, R. 2004. The emulation theory of representation: Motor control, imagery and perception. *Behavioral & Brain Sciences* 27: 377-442.

Hohwy, J. 2013. *The Predictive Mind*. Oxford: Oxford University Press.

Jebara, T. 2004. *Machine Learning: Discriminative and Generative*. New York: Springer.

Johnson-Laird, P. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. 2006. *How We Reason*. New York: Oxford University Press.

Kiefer, A. & J. Hohwy. 2018. Content and misrepresentation in hierarchical generative models. *Synthese* 195 (6): 2387-2415.

Kuvayev, L. & R. Sutton. 1996. Model-based reinforcement learning with an approximate, learned model. *Proceedings of the Ninth Yale Workshop on Adaptive and Learning Systems*, pp. 101-105.

Lake, B., T. Ullman, J. Tenenbaum & S. Gershman. 2017. Building machines that learn and think like people. *Behavioral & Brain Sciences* 40: e253.

Lee, S. W., S. Shimojo & J. O'Doherty. 2014. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* 81 (3): 687-699.

Magnani, L. & N. Nersessian. 2002. *Model-Based Reasoning: Science, Technology, Values*. New York: Springer.

Mnih, V. et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (7540): 529-533.

O'Brien, G. & J. Opie. 2004. Notes towards a structuralist theory of mental representation. In Clapin, Staines & Slezak, eds., *Representation in Mind: New Approaches to Mental Representation*. Oxford: Elsevier.

Quilty-Dunn, J. 2016. Iconicity and the format of perception. *Journal of Consciousness Studies* 23 (3-4): 255-263.

Quilty-Dunn, J. forthcoming. Perceptual pluralism. *Noûs*.

Ramsey, W. 2007. *Representation Reconsidered*. New York: Cambridge University Press.

Rusu, A., N. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu & R. Hadsell. 2016. Progressive neural networks. arXiv preprint.

Ryder, D. 2004. SINBAD neurosemantics: A theory of mental representation. *Mind & Language* 19: 211-240.

Shagrir, O. 2012. Structural representations and the brain. *British Journal for the Philosophy of Science* 63: 519-545.

Shea, N. 2014. Exploited isomorphism and structural representation. *Proceedings of the Aristotelian Society* 64: 123-144.

Shea, N. 2018. *Representation in Cognitive Science*. Oxford: Oxford University Press.

Suárez, M. 2003. Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science* 17: 225-244.

Sutton, R. 1991. Dyna, an integrated architecture for learning, planning and reacting. *SIGART Bulletin* 2 (4): 160-163.

Sutton, R. & A. Barto. Forthcoming. *Reinforcement Learning: An Introduction*. Second Edition.

Swoyer, C. 1991. Structural representation and surrogative reasoning. *Synthese* 87: 449-508.

Tenenbaum, J., C. Kemp, T. Griffiths & N. Goodman. 2011. How to grow a mind: Statistics, structure and abstraction. *Science* 331 (6022): 1279-1285.

Von Eckardt, B. 1993. *What is Cognitive Science?* Cambridge, MA: MIT Press.

Webb, T, & M. Graziano. 2015. The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology* 6.

Weisberg, M. 2013. *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.

Williams, D. & L. Colling. 2018. From symbols to icons: The return of resemblance in the cognitive neuroscience revolution. *Synthese* 195 (5): 1941-1967.

Wolpert, D., Z Ghahramani & M. Jordan. 1995. An internal model for sensorimotor integration. *Science* 269: 1880-1882.