Metadata of the chapter that will be visualized online

Chapter Title	Linking Models with Brain Measures	
Copyright Year	2024	
Copyright Holder	Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Love**
	Particle	
	Given Name	**Bradley C.**
	Suffix	
	Organization	University College London, The Alan Turing Institute
	Address	London, UK
	Email	b.love@ucl.ac.uk
Abstract	Linking models and brain measures offers a number of advantages over standard analyses. Models that have been evaluated on previous datasets can provide theoretical constraints and assist in integrating findings across studies. Model-based analyses can be more sensitive and allow for evaluation of hypotheses that would not otherwise be addressable. For example, a cognitive model that is informed from several behavioural studies could be used to examine how multiple cognitive processes unfold across time in the brain. Models can be linked to brain measures in a number of ways. The information flow and constraints can be from model to brain, brain to model, or reciprocal. Likewise, the linkage from model and brain can be univariate or multivariate, as in studies that relate patterns of brain activity with model states. Models have multiple aspects that can be related to different facets of brain activity. This is well illustrated by deep learning models that have multiple layers or representations that can be aligned with different brain regions. Model-based approaches offer a lens on brain data that is complementary to popular multivariate decoding and representational similarity analysis approaches. Indeed, these approaches can realise greater theoretical significance when situated within a model-based approach.	
Keywords (separated by "-")	Linking - Cognitive models - Multivariate measures of cognition	

Linking Models with Brain Measures

Bradley C. Love

Abstract Linking models and brain measures offers a number of advantages over standard analyses. Models that have been evaluated on previous datasets can provide theoretical constraints and assist in integrating findings across studies. Model-based analyses can be more sensitive and allow for evaluation of hypotheses that would not otherwise be addressable. For example, a cognitive model that is informed from several behavioural studies could be used to examine how multiple cognitive processes unfold across time in the brain. Models can be linked to brain measures in a number of ways. The information flow and constraints can be from model to brain, brain to model, or reciprocal. Likewise, the linkage from model and brain can be univariate or multivariate, as in studies that relate patterns of brain activity with model states. Models have multiple aspects that can be related to different facets of brain activity. This is well illustrated by deep learning models that have multiple layers or representations that can be aligned with different brain regions.

Model-based approaches offer a lens on brain data that is complementary to popular multivariate decoding and representational similarity analysis approaches. Indeed, these approaches can realise greater theoretical significance when situated within a model-based approach.

Keywords Linking · Cognitive models · Multivariate measures of cognition

1 Introduction

Psychology and neuroscience are concerned with theoretical concepts that cannot be directly measured. For example, theoretical concepts like recognition, familiarity, error, learning, replay, receptive field, fear, prejudice, value, and uncertainty need to

B. C. Love (✉)
University College London, The Alan Turing Institute, London, UK
e-mail: b.love@ucl.ac.uk

B. U. Forstmann, B. M. Turner (eds.), *An Introduction to Model-Based Cognitive Neuroscience*, https://doi.org/10.1007/978-3-031-45271-0_2

be operationalised. We cannot directly measure these concepts like we measure the temperature of a room with a thermometer or the length of a bolt with a ruler.

To further complicate matters, we are often interested in how processes unfold over time. For example, memory by definition involves processes that extend over time and involve generalisation or similarity structure. Likewise, decision-making processes, such as evidence accumulation for competing options, involve decision variables that change over time (Shadlen & Kiani, 2013). The dynamical nature of cognition is central in many accounts of behaviour (Busemeyer & Townsend, 1993; Tanenhaus et al., 1995; Wijeakumar et al., 2017).

To understand the brain basis of theoretical concepts in psychology, we need to measure these concepts and relate our measurements to the brain. Formal models offer one way forward. Models can be used to characterise cognitive processes in terms of the steps people carry out while performing a task. For example, drift-diffusion models (see chapter "Reinforcement Learning: Application to fMRI") characterise how evidence is accumulated over time for choice options (Ratcliff, 1978). Learning models characterise how knowledge is updated in light of corrective feedback, detailing the nature of error signals (Kruschke, 1992; Love et al., 2004). Cognitive models that have been rigorously evaluated are our best guess of how cognitive processes unfold. By fitting these models, such as to behavioural data, we can operationalise and quantify theoretical concepts of interest, akin to how a thermometer allows us to measure temperature.

One research goal in model-based neuroscience is to understand how abstract processes and representations detailed in cognitive models are instantiated in the brain (Forstmann et al., 2011; Palmeri et al., 2015; Turner et al., 2017). Additionally, as I will discuss, relating theoretical concepts to brain measures may also help advance our understanding of cognition by introducing additional constraints when fitting and selecting among candidate cognitive models. In effect, there can be a two-way street in which cognitive models help us to understand the brain and the brain helps us to develop and evaluate cognitive models.

Cognitive models can serve as the bridge between abstract theories and brain measures (Love, 2015). Model-based neuroscience offers the possibility of advancing our understanding along multiple levels of analysis. Linking models with brain measures also creates a number of exciting opportunities. As I will review, there are a number of cases in which brain imaging researchers could not have made an advance without a model-based analysis approach. In this chapter, I will consider several ways in which cognitive models can be related to brain measures and provide illustrative examples. As reviewed in Turner et al. (2017), cognitive models, which are concerned with behaviour, can be related to brain data in a number of ways, including (1) using the brain measures to constrain the cognitive model, (2) using the cognitive model to predict neural data, and (3) considering both the brain and behavioural data simultaneously. These approaches can be univariate or multivariate (i.e. patterns of brain activity are considered).

2 Some Functions of Models in Science

Models can play a number of constructive roles in psychology, neuroscience, and science more broadly. One function is simply organising one's ideas and making assumptions clear. Formal models require researchers to detail each step, which can reduce wiggle room relative to purely verbal theories. Whatever wiggle room is left (e.g. tuneable parameters) is made explicit.

As a consequence, what is predicted under different circumstances is made clear. Rather than debate what a theory predicts, a model can be simulated. For example, early work showing an advantage in processing category prototypes led researchers to believe that abstract prototypes were stored in memory, but subsequent work demonstrated that such effects were compatible with exemplar models that store no abstractions in memory (Medin & Schaffer, 1978). More recently, models have played a related role in the design and interpretation of fMRI (functional magnetic resonance imaging) studies of memory (Caplan & Madan, 2016; Nosofsky et al., 2012). Models can play a constructive role in directing empirical investigations.

Science often progresses by evaluating competing theoretical accounts. Models afford the possibility of model comparison in which competing accounts can be pitted against one another, and the model that performs best can be favoured. This approach is standard in mathematical psychology (Pitt et al., 2002) but can also be done in cognitive neuroscience. For example, Mack et al. (2013) formally evaluated whether the representations in an exemplar or prototype model best matched the BOLD (blood-oxygen-level-dependent) response and found that the exemplar model was more consistent (also see Stillesjö et al. (2019)). In such cases, brain data can help adjudicate between competing models when behavioural data alone cannot (Ditterich, 2010; Mack et al., 2013; Purcell et al., 2012). Recent work evaluating whether the hippocampus learns to associate objects and words incrementally or in an all-or-none fashion used a related approach that favoured the all-or-none account (Berens et al., 2018). Model comparison can even be done in cases in which behavioural data are not analysed. For example, recent work (Bobadilla-Suarez et al., 2019) asks what makes two brain states similar evaluating a number of basic accounts of similarity, such as Euclidean distance, Mahalanobis distance, Pearson correlation, etc., and found that the same similarity measures were operable across brain states but differed across tasks or stimuli.

Models can serve a powerful integrative role by linking seemingly disparate findings through common computational mechanisms. For example, a simple model of familiarity and recognition memory captured findings from both fMRI studies of visual categorisation and word list memory (Davis et al., 2014). In my own work, the same clustering approach for capturing behaviour in learning studies has been applied to a number of fMRI studies (Davis et al., 2012a, b; Inhoff et al., 2018; Mack et al., 2016, 2020). Applying the same model to multiple studies helps to theoretically integrate these empirical contributions, which is especially helpful when studies involve different paradigms and dependent measures. More recently, our clustering work (Mok & Love, 2019) has extended these same model mechanisms to offer an alternative explanation for place and grid cell responses in rodents and humans. This account makes novel predictions for how cell responses should change under different experimental conditions. In summary, cognitive models are useful tools for clarifying one's thinking, evaluating theoretical proposals, and, as will be discussed here, linking behaviour and brain.

3 Levels of Analysis

The aforementioned models can be considered cognitive models. These models are hypothesised to involve the same processes and representations as the human mind. Cognitive models reside at Marr's (1982) algorithmic level and are well placed to help explain how the brain implements higher-level computations (Love, 2015). As discussed below, the algorithmic level resides between higher-level considerations related to the description or goal of the overall computation and lower-level accounts of the computation's physical realisation, such as in the brain.

Marr's tripartite hierarchy (Marr, 1982) is perhaps the most well-known and influential organisation of levels in neuroscience. In brief, the computational level is the top level where the problem to be addressed is specified. Rather than detail the form of a potential solution, the computational level simply states the problem (i.e. the input-output mapping desired). For example, for object recognition, a computational-level account could involve naming various images under various conditions. The next level is the algorithmic level. As its name indicates, the algorithmic level is concerned with how the function specified at the computational level is computed (i.e. the processes and representations used). For example, if the computational-level task were to sort an array of numbers in ascending order, then the algorithmic level would specify a possible approach, such as bubble sort or quicksort. Different algorithms may solve the computational task in different ways, have different runtimes, etc., but they should all conform to the computational-level goal (e.g. correctly sort the array). Finally, the implementational level describes the physical substrate for the computation (e.g. the computer that executes quicksort).

The previous examples from computer science are apropos as Marr was clearly inspired by abstraction layers, a central concept in computer science (Wing, 2008). Note that Marr's top two levels, the computational and algorithmic, neatly map onto the top two levels in a common abstraction hierarchy in computing (Fig. 1). Abstraction layers in computing can contain finer-grain levels, including multiple levels describing the physical computing device. In contrast, Marr effectively lumped all of neuroscience into a single implementational level, which might partly explain why some neuroscientists find his hierarchy inadequate (Churchland et al., 1990).

Although Marr's scheme is highly influential, there are alternatives (Pylyshyn, 1984). Moreover, there is no reason to restrict to three levels. For example, there are a number of four-level schemes in cognitive science (Dawson, 2013; Newell, 1980, 1990; Sun, 2009). Indeed, Bechtel and Richardon's (1993) mechanistic

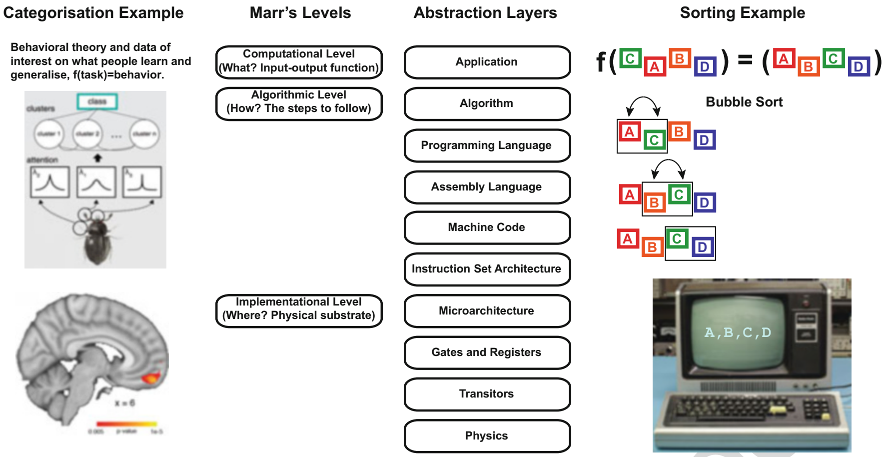

Fig. 1 Marr's levels compared to abstraction layers in computing with examples of each. Marr's levels are clearly influenced by abstraction layers in computer science, though Marr's levels are less fine grain, particularly for levels of interest to many neuroscientists. On the left, an example from category learning is shown in which an algorithmic model (Love et al., 2004) was fit to behaviour and its internal representations are used to interpret BOLD response (Mack et al., 2016). On the right, a sorting algorithm addressed the computational-level problem of sorting and was implemented by a digital computer. The abstraction layers in computing make clear that moving to a lower layer introduces additional detail (more information) about the computation whereas higher layers introduce abstract constructs that can be realised in multiple ways. (Figure and discussion from Love (2020a))

approach can be characterised as a "levels of mechanism" hierarchy in which there are not a fixed number of levels. For example, a car can be seen as mechanism consisting of interacting parts, such as an engine, drivetrain, steering wheel, brakes, etc. A component of a mechanism itself can be further decomposed into its own mechanism (e.g. braking system) and so forth with no limit except those imposed by particle physics.

For the present purposes, the important point is that cognitive models reside at an intermediary level that details the "how" of cognition. Given this placement, cognitive models can bridge between input-output descriptions of behaviour and brain implementation.

4 Other Types of Models Useful in Analysing Brain Data

In addition to using cognitive models, neuroscientists also use formal models as data analysis tools. For example, the generalised linear model (GLM) itself is a formal model that has assumptions and tuneable parameters that are fit to data. Of course, the GLM is not a model of how people process and represent information.

Returning to Marr's levels, it is clear that the GLM does not lie at the algorithmic level in understanding human cognition nor any other level. Instead, the GLM is an analysis tool.

Other examples of data analysis tools that are not cognitive models include dynamic causal modelling (Friston et al., 2003), techniques to measure the intrinsic or functional dimensionality of fMRI data (Ahlheim & Love, 2018), and multi-voxel pattern analysis (MVPA).

MVPA decoding approaches apply a machine classifier to "mind read" from the BOLD response whether a participant, for example, is viewing a house or a face (Cetron et al., 2019). Although these are not psychological models, they can be used to make interesting behavioural predictions. For example, participants tend to have faster response times for stimuli that are further from the classifier's decision bound, which indicates the classifier is more confident about its decision (Ritchie & Op de Beeck, 2019a). Decoding approaches can also be used to determine when people are engaging in replay (Lee et al., 2019; Momennejad et al., 2018; Shanahan et al., 2018; Xue, 2018).

There is a lot of room for creativity and innovation in using non-cognitive models, such as decoding procedures. For example, Shen et al. (2019) coupled a decoding approach with a deep convolutional network to visualise the image a person was viewing. Other methodological innovations include hyperalignment, which creates a common brain space for multiple participants to increase decoding performance (Haxby et al., 2011). Hyperalignment is successful because voxels do not exactly align across individuals' brains, but simple transformations to a common space can reveal commonalities across individuals.

The line distinguishing cognitive models and data analysis tools can be blurred at times. The distinction can depend on the intentions of the researcher using the model. Analogously, a Bayesian model can be taken as a computational-level theory of cognition (i.e. describing the behaviour that should occur under different circumstances with no recourse to the processes or representations that people use) or as algorithmic proposals of how people algebraically solve the task (Jones & Love, 2011a). For example, an algorithmic Bayesian model may predict response times depending on the nature of model updates, which are interpreted as mental operations, not computational-level descriptions. Making clear the nature of the model used is important because it determines how the model should be evaluated (Jones & Love, 2011b).

5 General Comparison of Model and Brain Data

A lot of early brain-inspired work in cognitive science was only loosely informed by findings in neuroscience. For example, the original parallel distributed processing (PDP) movement in the 1980s was motivated by the idea that brain computation is distributed across neurons and that cognitive models should reflect this observation (Rumelhart & McClelland, 1986). Notice this linkage between PDP models and the brain does not involve the fit of neural measures nor other formal coupling. Theoretical assertions of being brain-like or biologically plausible can be controversial in part because they are often underspecified whereas model selection procedures make claims and results clearer (Love, 2020a). The PDP models neglected many of the details of actual neurons, such as ion channels and spiking activity. Abstracting away details is not necessarily negative – in accord with Occam's razor, models should be as simple as possible while capturing the data of interest, which may or may not include the specifics of neurons. Again, model selection approaches make clear what data the scientist intends to explain.

The loose coupling of models and brain can be made somewhat more direct in cognitive models that attempt to simulate basic patterns of behaviour across different populations that vary in some key way, such as whether a group has a hippocampal lesion (Love & Gureckis, 2007; Nosofsky & Zaki, 1998). This basic approach is common and has been fruitful in exploring semantic processing impairments (Lambon Ralph et al., 2006; Tyler et al., 2000). Again, in these lines of work, cognitive modelling and analysis of brain data are happening separately from one another.

The relation between model simulations and brain measures can become quite rich. For example, recent work relates clustering mechanisms that have been used in concept learning to explain grid and place cell recordings in the rodent brain during navigation tasks (Mok & Love, 2019). In this case, the cognitive model is predicting how lower-level cell activity should vary with changes in task and environment. Although this work is theoretical and links cognitive models to the level of neurons, notice that this linkage does not involve exploiting any joint constraints in the data analysis. For example, the cognitive model is not being used to identify cell types by applying it to neural data. Instead, the model is being simulated and theoretically related to brain activity to help interpret and conceptualise findings.

In some sense, the entire emerging field of computational psychiatry falls under this heading of loosely connecting cognitive models to brain function. In computational psychiatry, cognitive models are routinely fit to behaviour, and fitted parameters for different populations (e.g. depressives vs. non-depressives) are compared (Adams et al., 2015; Blanco et al., 2013).

Certainly, work that provides a general conceptual link between brain and behaviour can be valuable. However, ideally, models would also be integrated into the data analysis. The remainder of this chapter focuses on incorporating cognitive models into the analysis of brain measures. Such model-based neuroscience approaches both theoretically relate cognitive models to the brain (as do the accounts reviewed in this section) and incorporate constraints across levels of analysis when evaluating models and brain data.

6 Cognitive Model as Integral Part of the Data Analysis

In a typical task fMRI (or EEG, MEG, etc.) analysis, experimental conditions are contrasted with one another. For example, one may contrast voxels that are more active for face than for house stimuli. The simplest model-based analyses replace the stimulus condition with some model measure (e.g. prediction error) that varies across trials (Daw et al., 2006). By entering this regressor (e.g. prediction error) from the cognitive model into the GLM, one can evaluate which voxels co-vary with the cognitive construct. As shown in Fig. 2, both the typical contrast approach and simple model-based analyses are univariate. Instead, standard MVPA start from a collection of voxels (multivariate) and aim to predict some experimental condition, such as whether the participant is viewing a house or a face. One innovation is to make the target of decoding a model measure, such as item familiarly according to a cognitive model (Mack et al., 2013). The four quadrants shown in Fig. 2 are not an exhaustive taxonomy of how to relate models to the BOLD response (for a more complete treatment, see Turner et al. (2017)).

Perhaps because it is relatively straightforward, the univariate model-based approach is most common in the field. Typically, a model is fit to behavioural data

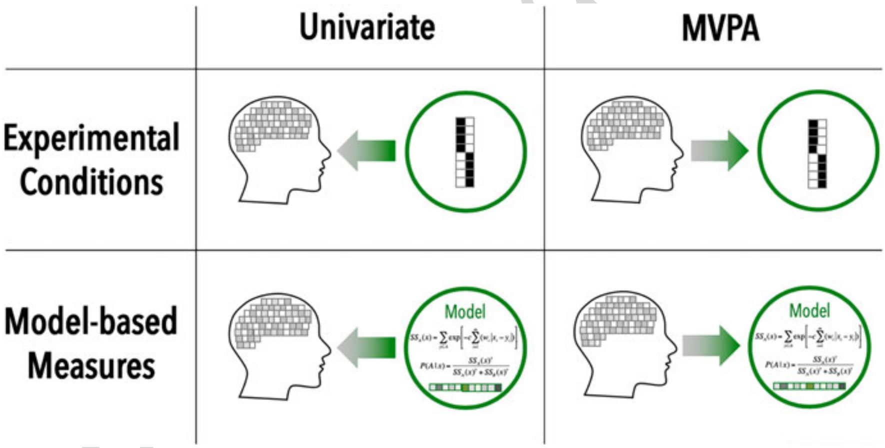

Fig. 2 The top row illustrates approaches that are not model-based in that they do not leverage a cognitive model of the task. For example, in the top-left panel, a standard analysis might identify voxels that are more active for faces than for house stimuli, whereas in the top-right panel, a decoder might try to classify whether the participant is viewing a house or a face stimulus on each trial. In the bottom row, a cognitive model is at the centre of the analysis. In the bottom-left panel, some measure from the cognitive model (which is usually fit to behavioural data), such as item familiarity, learning update, etc., is entered into the GLM. Such an analysis will identify voxels that show a similar activation profile to the model measure. In contrast, in the bottom-right quadrant, a classifier is applied to the brain to try to decode some internal measure from the cognitive model. In this case, models are favoured to the extent that their internal state is decodable (Mack et al., 2013). (Figure and discussion from Love (2020b))

and then used as a lens on the fMRI data. For example, an associative learning model was fit to behavioural data from a task where people formed impressions of various social groups through trial-by-trial feedback (Spiers et al., 2017). The fitted model provided a GLM trial-by-trial measure of valence or prejudice for each group, which tracked activity in the anterior temporal lobe in the model-based analysis. Model-based analysis was critical for capturing changes in memory *across* study trials.

In a category learning study (Davis et al., 2012a), a model-based analysis with a clustering model of learning was critical to capturing two time courses, one across trials and one within. This study examined the hippocampus' role in acquiring categories in which most items followed a rule but some items (exceptions) did not. A clustering model (Love et al., 2004) was fit to the behavioural data (i.e. the learning curves), and two model-based measures were entered into the GLM, one for recognition strength or familiarity and one for error correction or learning update. As shown in Fig. 3, the hippocampus tracked the model's recognition measure at stimulus presentation and the error measure at feedback presentation. Interestingly, a standard analysis contrasting exception and rule-following items found no significant difference – the cognitive model proved critical to capturing how hippocampal response changes over the course of study trials.

The same modelling approach can also be used to localise two simultaneous processes (by using two different model-based measures) within the same phase of a trial-to-draw distinction between the functions of anterior and posterior hippocampus (Davis et al., 2012b). Another way to scale up this basic univariate modelling approach is to adopt an encoder approach in which the fitted cognitive model provides a number of model-based regressors to enter into the GLM with the goal of explaining the most variance possible within brain regions of interest (van Gerven, 2017). In the encoding approach, rather than trying to identify voxels that significantly regress on some specific model-based measure (e.g. prediction error), the goal is for multiple model measures to capture the most overall variance possible in the GLM.

Another model-based work (Kragel et al., 2015; Palmeri et al., 2015) reverses the flow of information to incorporate brain measures directly into the operation of the model to better predict behaviour. For example, Kragel et al. (2015) used a variant of the context maintenance and retrieval (CMR) model of free recall (Polyn et al., 2009) that took signals from the medial temporal lobe (MTL) to determine whether contextual reactivation was successful at each potential recall event. The model that incorporated the BOLD input performed better than a baseline model in predicting behaviour. Another example of this approach is replacing parameters in decision models, such as in drift-diffusion model (Ratcliff, 1978) and variants (Usher & McClelland, 2001) with neural recordings from regions thought to implement the functions of those parameters (Palmeri et al., 2015; Purcell et al., 2010).

Rather than linking from model to brain or brain to model, joint modelling approaches (Turner et al., 2019a, b) simultaneously model the mutual constraints between behavioural and brain measures through an intermediary cognitive model. This approach can deal with multiple brain measures (e.g. fMRI and EEG) and can make predictions about missing measures based on covariance with the observed

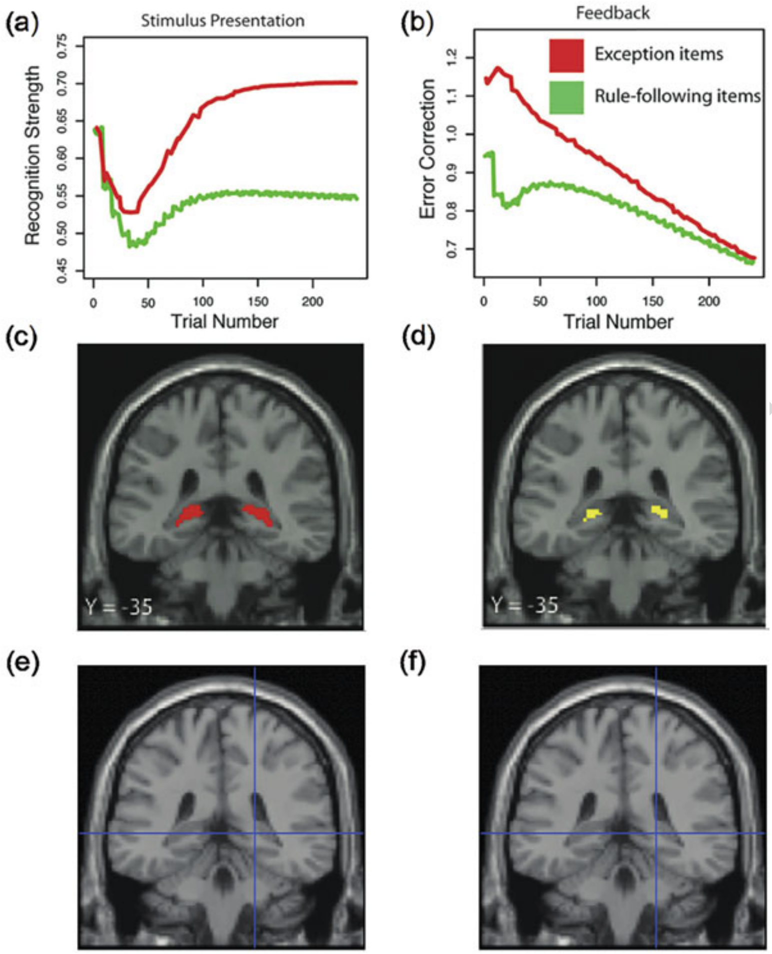

Fig. 3 Panels a and b show model-based regressors for a measure of recognition strength (i.e. familiarity) and error correction (i.e. learning update). These model-based regressors track hippocampal activity at the stimulus presentation and feedback phases of trials, respectively (Davis et al., 2012a). In contrast, a standard contrast of exception > rule-following items (panels **e** and **f**) results in no statistically significant voxels, because this contrast does not track the time course of hippocampal activitys

measures. This approach can be quite powerful and useful in practice. For example, one could collect behavioural data from a number of participants and more costly neural recordings from only a subset of participants and leverage the constraints across measures and participants through hierarchal Bayesian modelling.

There are a number of other creative ways to link cognitive models to BOLD response. One way is to link a key event, as indexed by the cognitive model, to an operation in the brain. For example, a recent study finds that prediction errors during study are predictive of later replay events (Momennejad et al., 2018). In other work, a Bayesian model determined the probability that an item would be remembered, which correlated with hippocampal activity during encoding (Gluth et al., 2015).

Finally, a cognitive model's fitted parameters can be related to the BOLD response instead of a trial-by-trial measure from the model. During category learning, models (Love et al., 2004; Nosofsky, 1986) predict that goal-relevant aspects of the stimuli will receive greater weight or attention. A recent study found that the learned attentional weights from category learning models fit to behaviour were predictive of how well those stimulus aspects could be decoded from the BOLD response (Braunlich & Love, 2019). Relatedly, in a study exploring vmPFC (ventromedial prefrontal cortex)-hippocampal interactions during concept learning (Mack et al., 2020), the pattern of goal-directed representation compression in vmPFC paralleled the attention weights from a model fitted to behaviour.

7 Individual Differences

Both behavioural and brain measures, such as fMRI's BOLD response, tend to be very noisy both within and across individuals. Somewhat surprisingly, cognitive models that are fit to individual's behaviour can be used to understand individual differences in brain response. For example, in studies of category learning, individuals learn to attend to relevant stimulus dimensions that discriminate between the category responses (Kruschke, 1992; Love et al., 2004; Nosofsky, 1986). According to the fits of cognitive models, individuals' attentional strategies differ slightly from one another, which affects how attended each stimulus dimension is. Interestingly, these individual differences in attention weights arising from fitting behaviour can also be observed in brain response – stimulus aspects that are more attended by an individual are easier to decode in visual areas using MVPA (i.e. mind reading) on the fMRI BOLD response (Braunlich & Love, 2019). Relatedly, compression signals found in the ventromedial prefrontal cortex (vmPFC) thought to relate to attentional allocation and also relate to individual differences in attentional weighting over the course of learning. A final example comes from the neuroeconomics literature from a task patterned after shopping on Amazon. Participants' willingness to update their beliefs in the face of Amazon reviews was modelled by a Bayesian model fit to behaviour with the tendency of an individual to update, correlating with overall activity in the dorsomedial prefrontal cortex (De Martino et al., 2017).

In the aforementioned analyses, estimates for individuals were independent from another in that individuals were not linked during the analysis. An alternative approach, such as in Bayesian hierarchal modelling, is to assume that individuals belong to a common family such that estimates of individual inform the estimates for others. When data are noisy, hierarchal approaches that link estimates may offer advantages and have been used successfully in modelling individual differences in cognitive control (Molloy et al., 2019). When using an independent or hierarchal approach, the conclusion that cognitive models can reflect a reality at both the behavioural and neural levels for individual participants is exciting and demonstrates how modelling can extract fine-grain information.

8 Models Can Uncover Useful Latent States

Models can be useful in inferring latent states that can help explain behaviour and its brain basis. One example of latent variables are the clusters in the aforementioned learning models (Anderson, 1991; Love et al., 2004) which detail how related items are stored together in memory (Mack et al., 2018). Models operationalise these hypothesised representational structures, which can be useful in analysing BOLD response.

Inferring latent state is more complex when researchers aim to characterise complex mental operations that unfold through time (Wijeakumar et al., 2017). One popular approach is to use hidden Markov models (HMMs) to infer what operations people are currently undertaking and using this characterisation to interpret the BOLD response (Anderson et al., 2018; Tubridy et al., 2018).

The importance of inferring latent state is also becoming appreciated in related fields, such as reinforcement learning (Niv, 2019). Many of the same conceptual issues and brain systems are implicated in these tasks as in goal-directed concept learning. For example, strategic exploration relies on hippocampal-prefrontal cooperation (Wang & Voss, 2014) as is found during memory tasks (Mack et al., 2020).

9 Comparing Model and Brain Representations

In addition to MVPA decoding, multivariate pattern analysis can be used to compare proposed (e.g. model) representations and voxel representations (Haxby, 2001). This pattern comparison analysis is popularly known as representational similarity analysis (RSA) (Dimsdale-Zucker & Ranganath, 2018). RSA correlates two similarity matrices, one from the cognitive model and one from the brain, to assess how well the two similarity spaces align. RSA can be used as confirmatory evidence that a model provides the correct representational account of a brain region or in an exploratory fashion such as in a whole-brain searchlight analysis. One application of RSA is to compare proposed memory representations acquired by models of concept learning to brain regions thought to implement those functions (Mack et al., 2013; Ritchie & Op de Beeck, 2019b). For example, RSA analyses found that hippocampal representations of objects (see Fig. 4) are modulated by changes in the task goal (Mack et al., 2016).

For an RSA to be model-based, one of the similarity matrices should be generated by a cognitive model. RSA can involve the evaluation of several cognitive models. A variety of models can be considered, and the model whose representations best align with the brain can be favoured (Ritchie & Op de Beeck, 2019b). However,

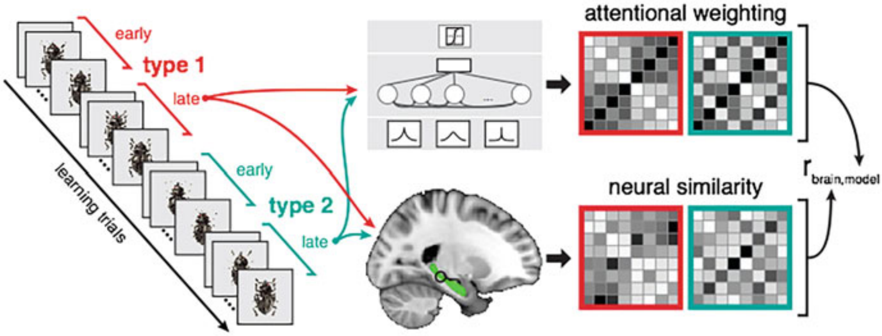

Fig. 4 Representation similarity analysis (RSA) can be used to compare a cognitive model's representations to those of the brain. In this example (Mack et al., 2016), a cognitive model was fit to behaviour for different learning problems (shown in red and teal). For each problem, the cognitive model was used to calculate a similarity matrix for the stimulus items. Similarity matrices were also calculated by comparing voxel activity for the stimulus items. In the left anterior hippocampus, the similarity patterns predicted by the model and those observed in the brain agreed

not all RSAs are model-based and the dividing line can be blurry. For example, technically, finding that hippocampus CA1 codes distance to a goal (Spiers et al., 2018) is not model-based (because distance is specified by the task), whereas coding distance to some model quantity, such as distance to a category prototype (Seger et al., 2015), is model-based (because the prototype is specified by the fitted cognitive model). For a model-based analysis to be useful, it should add something beyond a standard analysis. Ideally, a model-based analysis would improve both data fit and our understanding of the domain. For example, a model may largely code distance to goal but diverge in informative ways under certain circumstances that could be empirically verified and in turn deepen our understanding of the domain.

Certainly, univariate analyses can be rigorous, interesting, and motivated but not model-based. The same is true in RSA. For example, a recent study (Martin et al., 2018) used similarity matrices designed to capture perceptual or conceptual similarity to hone in on the function of perirhinal cortex and other regions. This work is exciting and valuable, but because the similarity matrices were derived from human ratings rather than generated by a model of perceptual or conceptual processing, the analysis is not model-based.

Although RSA is popular and powerful, it is not entirely clear what advantages it offers over general statistical approaches such as canonical correlation analysis (CCA) or related techniques such as partial least squares (PLS). CCA maximises the correlation between two sets of multivariate measurements. For example, one set of measures could be on the brain side, such as a collection of voxels or the time course for an individual voxel, and the other set of measures could be from a cognitive model, a set of experiment ratings, etc. Although CCA has been used in imaging analysis and software tools exist (Bilenko & Gallant, 2016), it is not as popular as RSA at the present time, though that could change as CCA seems to offer a number of advantages (e.g. it infers weights for the individual measures in the two domains, takes the reliability of measures into account, etc.) and no disadvantages that I can discern. It is also preferred over RSA for related problems, such as comparing representations from deep learning networks (Morcos et al., 2018).

10 Multiple Levels of Representation

The advent of deep learning has opened a number of possibilities in model-based neuroscience. Deep learning models are the descendants of connectionist models that were prominent in psychology in the 1980s (Rumelhart & McClelland, 1986). Like those earlier models, the weights in deep learning models are typically trained end-to-end through gradient descent procedures. Through architectural innovations, such as multiple convolutional and pooling layers, these networks display abilities that eclipse their predictors and excel at computer vision benchmarks (Krizhevsky et al., 2012). Despite being developed for engineering purposes, these models provide leading accounts of computation along the human and monkey ventral stream (Guclu & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kubilius et al., 2018; Yamins & DiCarlo, 2016). They have also been useful for exploring ideas about the nature of neural code (Guest & Love, 2017). Because deep learning models can take photographic stimuli as input, they open a number of opportunities for researchers, such as using these networks to derive stimuli that should best drive the response of a brain region (Bashivan et al., 2019).

One positive aspect of these models is that they contain multiple levels of representation (see Fig. 5). Each layer of the model takes as input the output of the previous layer and transforms it, such that the initial input is a photograph and the final output is an object recognition decision. At each step in this transformation, the representations can be compared to the activity patterns in brain regions. One common finding is that the early and late layers in models tend to correspond to early and late regions along the visual ventral stream (Guclu & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kubilius et al., 2018; Yamins & DiCarlo, 2016). Model representations can be related to brain response using either RSA or encoder approaches. Although these models have been successful in accounting for object recognition and activity along the ventral stream, one future challenge is to incorporate additional processes, such as top-down, goal-directed attention (Lindsay & Miller, 2018; Roads & Love, 2019).

11 Conclusions

Adopting a model-based approach to analysing brain measures offers a number of advantages. In some cases, one can evaluate hypotheses that otherwise would not be possible with a standard analysis approach. Models, which formalise related

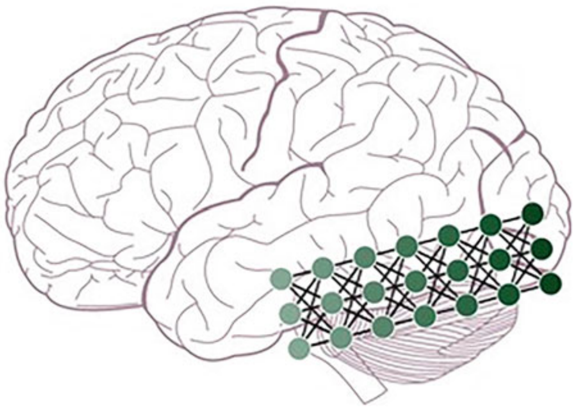

Fig. 5 Deep learning models contain multiple layers or processing stages that transform the stimulus. This enables evaluation of hypotheses that span brain regions, such as that the levels of object recognition deep networks correspond to stages along the ventral visual stream. (Image from Guest and Love (2017))

theories, offer the hope that results will be theoretically grounded. As related models are applied across data sets, models may promote a more systematic and cohesive science. Cognitive models are well positioned to integrate findings across levels of analysis (Love, 2015).

I have reviewed a number of ways to relate cognitive models to brain response. Possibilities include fitting models to behaviour and incorporating derived trial-by-trial measures into the GLM, model decoding approaches (Mack et al., 2013), using brain response to drive the behavioural predictions of the model, joint modelling to simultaneously address brain and behavioural measures, and comparing model representations and brain response. Which approach is suitable is largely a function of the study's design and the researcher's aims.

Opportunities and choices in conducting model-based analysis of brain data are rapidly increasing. It is an exciting time as there is latitude to be creative whether one is applying an existing technique or developing a novel analysis approach to address a new challenge. Although flexibility in inference can lead to false positives, model-based analyses can provide additional constraints by linking measures and multiple datasets. Model-based approaches can offer more stringent tests of theories and the possibility of comparing competing models. As open science initiatives and data repositories, such as OpenNeuro, make more datasets publicly available, the importance of model-based approaches, especially those that link multiple datasets, will only increase. Against this backdrop, modellers should do their part by making their code and details of their analyses publicly available through hosting and version control services such as GitHub.

One key question to consider is why do model-based analyses work? Models are not magical nor guaranteed to be helpful, so why are there so many cases in which model-based analyses succeed in pulling more from the data than would be possible through a standard analysis? The answer is that models have the ability to incorporate constraints that are outside the immediate study. In my own work, models are developed over years and honed while being applied to multiple behavioural and fMRI datasets. In this sense, the models have a reality and value outside their immediate application, which is critical because a model-based analysis is only as credible as the model used.

Acknowledgements This work was supported by the NIH Grant 1P01HD080679, Wellcome Trust Investigator Award WT106931MA, and Royal Society Wolfson Fellowship 183029 to B.C.L. Although mostly original, this paper draws on some previously published work (Love, 2020a, b; Turner et al., 2017). Thanks to Sebastian Bobadilla-Suarez for helpful comments on a previous draft.

Conflict of Interest Nothing declared.

Questions for Consideration

Model-based analyses can offer additional theoretical constraints but can also introduce degrees of freedom when choosing which model-based analysis to conduct. How should one choose which model-based analysis to conduct?

How much should we demand of researchers in terms of verifying their models before conducting a model-based analysis given that the analysis is only as good as the model used?

Will behavioural studies be increasingly valued as one avenue to verify models for model-based neuroscience?

The motivation for a model-based analysis can involve more than the model itself to include the bridge theory that links model components to brain regions. How does one choose between this focused, top-down approach to model application and a bottom-up, data-driven approach?

Models can be specified at multiple levels of abstraction (see "levels of mechanism" discussion). Why is it rare to have multiple models for the same task that differ in their level of abstraction?

Further Reading

- Love, B. C. (2020a). Levels of biological plausibility. *Philosophical Transactions of the Royal Society B*. https://doi.org/10.1098/rstb.2019.0632
- Love, B. C. (2020b). Model-based fMRI analysis of memory. *Current Opinion in Behavioral Sciences, 32*, 88–93. https://doi.org/10.1016/j.cobeha.2020.02.012
- Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology, 76*, 65–79.
- Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint models of neural and behavioral data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-03688-1

References

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, jnnp-2015-310737. https://doi.org/10.1136/jnnp-2015-310737

Ahlheim, C., & Love, B. C. (2018). Estimating the functional dimensionality of neural representations. *NeuroImage, 179*, 51–62. https://doi.org/10.1016/j.neuroimage.2018.06.015

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Anderson, J. R., Borst, J. P., Fincham, J. M., Ghuman, A. S., Tenison, C., & Zhang, Q. (2018). The common time course of memory processes revealed. *Psychological Science, 29*(9), 1463–1474. https://doi.org/10.1177/0956797618774526

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science, 364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton University Press.

Berens, S. C., Horst, J. S., & Bird, C. M. (2018). Cross-situational learning is supported by propose-but-verify hypothesis testing. *Current Biology, 28*(7), 1132–1136.e5. https://doi.org/10.1016/j.cub.2018.02.042

Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in Neuroinformatics, 10*. https://doi.org/10.3389/fninf.2016.00049

Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G., & Love, B. C. (2013). The influence of depression symptoms on exploratory decision-making. *Cognition, 129*(3), 563–568. https://doi.org/10.1016/j.cognition.2013.08.018

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, B. C. (2019). Measures of neural similarity. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-019-00068-5

Braunlich, K., & Love, B. C. (2019). Occipitotemporal representations reflect individual differences in conceptual knowledge. *Journal of Experimental Psychology: General, 148*(7), 1192–1203. https://doi.org/10.1037/xge0000501

Busemeyer, J. R., & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision-making in an uncertain environment. *Psychological Review, 100*, 432–459.

Caplan, J. B., & Madan, C. R. (2016). Word imageability enhances association-memory by increasing hippocampal engagement. *Journal of Cognitive Neuroscience, 28*(10), 1522–1538. https://doi.org/10.1162/jocn_a_00992

Cetron, J. S., Connolly, A. C., Diamond, S. G., May, V. V., Haxby, J. V., & Kraemer, D. J. M. (2019). Decoding individual differences in STEM learning from functional MRI data. *Nature Communications, 10*(1), 2027. https://doi.org/10.1038/s41467-019-10053-y

Churchland, P. S., Koch, C., & Sejnowski, T. J. (1990). What is computational neuroscience? In *Computational neuroscience* (pp. 46–55). MIT Press.

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based FMRI reveals specialized representations for surprising category members. *Cerebral Cortex, 22*(2), 260–273. https://doi.org/10.1093/cercor/bhr036

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 821–839. https://doi.org/10.1037/a0027865

Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience, 34*(22), 7472–7484. https://doi.org/10.1523/JNEUROSCI.3376-13.2014

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876–879.

Dawson, M. R. W. (2013). *Mind, body, world: Foundations of cognitive science*. Athabasca University Press.

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *The Journal of Neuroscience, 37*(25), 6066–6074. https://doi.org/10.1523/JNEUROSCI.3880-16.2017

Dimsdale-Zucker, H. R., & Ranganath, C. (2018). Representational similarity analyses. In *Handbook of behavioral neuroscience* (Vol. 28, pp. 509–525). Elsevier. https://doi.org/10.1016/B978-0-12-812028-6.00027-6

Ditterich, J. (2010). A comparison between mechanisms of multi-alternative perceptual decision making: Ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Frontiers in Neuroscience, 4*. https://doi.org/10.3389/fnins.2010.00184

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Sciences, 15*(6), 272–279. https://doi.org/10.1016/j.tics.2011.04.002

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage, 19*(4), 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7

Gluth, S., Sommer, T., Rieskamp, J., & Büchel, C. (2015). Effective connectivity between hippocampus and ventromedial prefrontal cortex controls preferential choices from memory. *Neuron, 86*(4), 1078–1090. https://doi.org/10.1016/j.neuron.2015.04.023

Guclu, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Guest, O., & Love, B. C. (2017). What the success of brain imaging implies about the neural code. *eLife, 6*, e21397. https://doi.org/10.7554/eLife.21397

Haxby, J. V. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Inhoff, M. C., Libby, L. A., Noguchi, T., Love, B. C., & Ranganath, C. (2018). Dynamic integration of conceptual information during learning. *PLoS One, 13*(11), e0207357. https://doi.org/10.1371/journal.pone.0207357

Jones, M., & Love, B. C. (2011a). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences, 34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134. discussion 188–231.

Jones, M., & Love, B. C. (2011b). Pinning down the theoretical commitments of Bayesian cognitive models. *Behavioral and Brain Sciences, 34*(4), 215–231. https://doi.org/10.1017/S0140525X11001439

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kragel, J. E., Morton, N. W., & Polyn, S. M. (2015). Neural activity in the medial temporal lobe reveals the fidelity of mental time travel. *Journal of Neuroscience, 35*(7), 2914–2926. https://doi.org/10.1523/JNEUROSCI.3378-14.2015

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc..

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the neural mechanisms of core object recognition [Preprint]. *Neuroscience*. https://doi.org/10.1101/408385

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2006). Neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain, 130*(4), 1127–1137. https://doi.org/10.1093/brain/awm025

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2019). Differential representations of perceived and retrieved visual information in hippocampus and cortex. *Cerebral Cortex, 29*(10), 4452–4461. https://doi.org/10.1093/cercor/bhy325

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife, 7*, e38105. https://doi.org/10.7554/eLife.38105

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science, 7*(2), 230–242. https://doi.org/10.1111/tops.12131

Love, B. C. (2020a). Levels of biological plausibility. *Philosophical Transactions of the Royal Society B*. https://doi.org/10.1098/rstb.2019.0632

Love, B. C. (2020b). Model-based fMRI analysis of memory. *Current Opinion in Behavioral Sciences, 32*, 88–93. https://doi.org/10.1016/j.cobeha.2020.02.012

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience, 7*(2), 90–108.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology, 23*, 2023–2027.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences, 113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters, 680*, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications, 11*(1), 46. https://doi.org/10.1038/s41467-019-13930-8

Marr, D. (1982). *Vision*. W. H. Freeman.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife, 7*, e31873. https://doi.org/10.7554/eLife.31873

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

Mok, R. M., & Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature Communications, 10*(1), 5685. https://doi.org/10.1038/s41467-019-13760-8

Molloy, M. F., Bahg, G., Lu, Z.-L., & Turner, B. M. (2019). Individual differences in the neural dynamics of response inhibition. *Journal of Cognitive Neuroscience, 31*(12), 1976–1996. https://doi.org/10.1162/jocn_a_01458

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife, 7*, e32548. https://doi.org/10.7554/eLife.32548

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 5727–5736). Curran Associates, Inc. http://papers.nips.cc/paper/7815-insights-on-representational-similarity-in-neural-networks-with-canonical-correlation.pdf

Newell, A. (1980). Physical symbol systems*. *Cognitive Science, 4*(2), 135–183. https://doi.org/10.1207/s15516709cog0402_2

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience, 22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Nosofsky, R. M. (1986). Attention, similairty, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M., & Zaki, S. F. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science, 9*, 247–255.

Nosofsky, R. M., Little, D. R., & James, T. W. (2012). Activation in the neural network responsible for categorization and recognition reflects parameter changes. *Proceedings of the National Academy of Sciences of the United States of America, 109*(1), 333–338. https://doi.org/10.1073/pnas.1111304109

Palmeri, T. J., Schall, J. D., & Logan, G. D. (2015). In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Neurocognitive modeling of perceptual decision making* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.15

Pitt, M. A., Myung, I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*(1), 129–156. https://doi.org/10.1037/a0014420

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review, 117*(4), 1113–1143. https://doi.org/10.1037/a0020311

Purcell, B. A., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2012). From salience to saccades: Multiple-alternative gated stochastic accumulator model of visual search. *Journal of Neuroscience, 32*(10), 3433–3446. https://doi.org/10.1523/JNEUROSCI.4622-11.2012

Pylyshyn, Z. W. (1984). *Computation and cognition. Toward a foundation for cognitive science*. MIT Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ritchie, J. B., & Op de Beeck, H. (2019a). Using neural distance to predict reaction time for categorizing the animacy, shape, and abstract properties of objects. *Scientific Reports, 9*(1), 13201. https://doi.org/10.1038/s41598-019-49732-7

Ritchie, J. B., & Op de Beeck, H. (2019b). A varying role for abstraction in models of category learning constructed from neural representations in early visual cortex. *Journal of Cognitive Neuroscience, 31*(1), 155–173. https://doi.org/10.1162/jocn_a_01339

Roads, B. D., & Love, B. C. (2019). Learning as the unsupervised alignment of conceptual systems. *ArXiv:1906.09012 [Cs, Stat]*. http://arxiv.org/abs/1906.09012

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition; Volume 1: Foundations*. MIT Press.

Seger, C. A., Braunlich, K., Wehe, H. S., & Liu, Z. (2015). Generalization in category learning: The roles of representational and decisional uncertainty. *Journal of Neuroscience, 35*(23), 8802–8812. https://doi.org/10.1523/JNEUROSCI.0654-15.2015

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron, 80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Shanahan, L. K., Gjorgieva, E., Paller, K. A., Kahnt, T., & Gottfried, J. A. (2018). Odor-evoked category reactivation in human ventromedial prefrontal cortex during sleep promotes memory consolidation. *eLife, 7*, e39681. https://doi.org/10.7554/eLife.39681

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology, 15*(1), e1006633. https://doi.org/10.1371/journal.pcbi.1006633

Spiers, H. J., Love, B. C., Le Pelley, M. E., Gibb, C. E., & Murphy, R. A. (2017). Anterior temporal lobe tracks the formation of prejudice. *Journal of Cognitive Neuroscience, 29*(3), 530–544. https://doi.org/10.1162/jocn_a_01056

Spiers, H. J., Olafsdottir, H. F., & Lever, C. (2018). Hippocampal CA1 activity correlated with the distance to the goal and navigation performance. *Hippocampus, 28*(9), 644–658. https://doi.org/10.1002/hipo.22813

Stillesjö, S., Nyberg, L., & Wirebring, L. K. (2019). Building memory representations for exemplar-based judgment: A role for ventral precuneus. *Frontiers in Human Neuroscience, 13*, 228. https://doi.org/10.3389/fnhum.2019.00228

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research, 10*(2), 124–140. https://doi.org/10.1016/j.cogsys.2008.07.002

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

Tubridy, S., Halpern, D., Davachi, L., & Gureckis, T. M. (2018). A neurocognitive model for predicting the fate of individual memories [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/7r3jp

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology, 76*, 65–79.

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019a). *Joint models of neural and behavioral data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-03688-1

Turner, B. M., Palestro, J. J., Miletić, S., & Forstmann, B. U. (2019b). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neuroscience & Biobehavioral Reviews, 102*, 327–336. https://doi.org/10.1016/j.neubiorev.2019.04.018

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75*(2), 195–231. https://doi.org/10.1006/brln.2000.2353

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

van Gerven, M. A. J. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology, 76*, 172–183. https://doi.org/10.1016/j.jmp.2016.06.009

Wang, J. X., & Voss, J. L. (2014). Brain networks for exploration decisions utilizing distinct modeled information types during contextual learning. *Neuron, 82*(5), 1171–1182. https://doi.org/10.1016/j.neuron.2014.04.028

Wijeakumar, S., Ambrose, J. P., Spencer, J. P., & Curtu, R. (2017). Model-based functional neuroimaging using dynamic neural fields: An integrative cognitive neuroscience approach. *Journal of Mathematical Psychology, 76*, 212–235. https://doi.org/10.1016/j.jmp.2016.11.002

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366*(1881), 3717–3725. https://doi.org/10.1098/rsta.2008.0118

Xue, G. (2018). The neural representations underlying human episodic memory. *Trends in Cognitive Sciences, 22*(6), 544–561. https://doi.org/10.1016/j.tics.2018.03.004

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365. https://doi.org/10.1038/nn.4244

AUTHOR QUERY

AQ1. Please provide caption for part figures “a–d” in Fig. 3.