# Different ways of linking behavioral and neural data via computational cognitive models

**Abbreviated title:** Linking behavioral and neural data

**Authors:** Gilles de Hollander [1,2,Ŧ], Birte U. Forstmann[1,2], Scott D. Brown[3]

**Affiliations:**

1. Amsterdam Brain & Cognition Center, University of Amsterdam, the Netherlands
2. Department of Psychology, University of Amsterdam, The Netherlands
3. School of Psychology, University of Newcastle, Australia

**Ŧ Corresponding author**

Gilles de Hollander
Department of Psychology
University of Amsterdam
Weesperplein 4, 1018XA Amsterdam
the Netherland+31 205256281
gilles.de.hollander@gmail.com

**Number of pages:** 32
**Numbers of figures:** 1
**Number of tables:** 1
**Number of words:** 5962

**Conflict of Interest:** The authors declare no competing financial interest.

**Acknowledgements**

This work was partially supported by a starter grant from the European Research Council (ERC) and a Vidi grant from The Netherlands Organisation for Scientific Research (NWO) to BUF, as well as the Australian Research Council's fellowship to SDB (FT120100244)

## Abstract

Cognitive neuroscientists sometimes apply formal models to investigate how the brain implements cognitive processes. These models describe behavioral data in terms of underlying, latent, variables linked to hypothesized cognitive processes. A goal of model-based cognitive neuroscience is to link these variables to brain measurements, which can advance progress in both cognitive and neuroscientific research. However, the details and the philosophical approach for this linking problem can vary greatly. We propose a continuum of approaches which differ in the degree of tight, quantitative, and explicit hypothesizing. We describe this continuum using four points along it, which we dub ``qualitative structural'', ``qualitative predictive'', ``quantitative predictive'', and ``single model'' linking approaches. We further illustrate by providing examples from three research fields (decision making, reinforcement learning, and symbolic reasoning) for the different linking approaches.

## Introduction

In recent years, cognitive neuroscientists have applied formal, computational cognitive models to more effectively understand how the brain implements cognitive processes such as decision making, reinforcement learning, and symbolic reasoning. Such formal cognitive models can decompose effects in behavioral data by description in terms of underlying latent cognitive processes and associated variables. ``Model-based cognitive neuroscience'' links these variables to brain measurements. This approach can, on the one hand, constrain the development of cognitive models, while, on the other hand, also refine models that explain how cognitive processes are implemented in the brain (1). Linking brain measurements to psychological constructs has been conceptualized as identifying a *bridge locus*: to link some mental capacity to a neural substrate (2; 3). A researcher can identify *bridge loci* by empirically testing probable *linking hypotheses*. An example of a linking hypothesis is that the ventral striatum represents a how much reward a subject expects during a task.

The scope of this paper is limited to the neural linking of computational cognitive models and excludes (much more common) conceptual verbal theories of cognition. A main strength of computational models of cognition over verbal theories is that they force the modeler to be explicit and precise in their assumptions about cognition. This reduces the potential for miscommunication and misunderstanding of what a cognitive theory entails and reduces the potential for vague statements that are hard to test empirically (4-6). Additionally, because of their quantitative nature, computational cognitive models offer the possibility of assigning hard numbers to abstract cognitive concepts like “response caution” or “learning rate”. These numbers allow the integration of cognitive theory with quantitative neural data in a statistical framework. Ultimately, we believe that this quantitative, statistical approach can bring us much tighter integration between the cognitive and neural domain than verbal theories, supporting more stringent tests of the theories and of the links between neural and behavioral data.

David Marr (7) famously subdivided the problem of understanding how the brain works into three levels: 1) a computational level that describes what computational problem a brain aims to solve in a given context, 2) an algorithmic level that describes how the problem can be solved, and 3) an implementational level that describes how this algorithm can physically be performed. Linking cognitive models to neural data can inform theories at all three levels.

For example, at the algorithmic level, cognitive models of speeded decision making make clear predictions about how subjects can lower the distance an evidence accumulator has to travel from the start of a trial to the end. However, for many models, it is not possible to investigate whether this is achieved by increasing the starting point or the finishing threshold of the accumulator, with only behavioral data. Clearly, neural data can help to distinguish between these different algorithms and needs to carefully be related to the cognitive models that are successful in explaining behavior (5; 8).

Similarly, more elaborate explanations at the implementational level are only possible with a firm understanding of what problem the brain is actually solving and what possible strategies are. This point is made again in Marr's original proposal of the three levels, and also eloquently put in the following analogy of David A. Robinson (9): "Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feather: it just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense". We believe the linking of cognitive models, explaining how computational problems the brain is faced with can be solved, to neural data, rooted in the physical substrate of the algorithms.

Another advantage of linking cognitive models to neural data might be the sheer wealth of additional information that neural data can provide in comparison to behavioral data. By any measure, the amount of information in behavioral data is extremely limited. Because many behavioral experiments provide not much more than choices and reaction times, literally all the data of a behavioral experiment can usually be summarized in a few hundred (only choice) up to a few thousand bytes (also reaction times). Compare that to ultra-high resolution fMRI data from 7 Tesla MR scanners, which can easily occupy a few billion bytes per subject. Of course the picture is more complicated than this: the neural data is much more ambiguous. However, recent efforts in both sequential sampling models as well as models of value-based learning have taught us that to reliably estimate the parameters of more complicated cognitive models and dissociate between different versions of them, the amount of information of most behavioral datasets is very limited (10-12). Thus, even disregarding the conceptual benefits, cognitive modelers should welcome the practical benefits arising from the wealth of extra information in neural data, as it provides an opportunity to develop richer models of cognition than has been possible so far.

But how do we link cognitive models to functional brain measurements most effectively? In the past decade, parameters of formal cognitive models have been linked to many measures of neural activity, such as EEG, fMRI, and single-cell recordings. These studies have employed wildly varying approaches, connecting variability in behavior and brain measurement at the level of subjects, conditions, and even trials. In some studies, cognitive models are used to set up testable hypotheses about brain activity. In other studies, cognitive model parameters are directly correlated against measurement models of neural data, after both models have been fit to their respective data domain. Some studies make a single model of both brain and behavior and try to predict both at the same time.

< INSERT TABLE 1 SOMEWHERE HERE>

In this review we aim to provide a particular taxonomy of possible methods of linking neural data to cognitive models. We think this taxonomy is useful to describe the work that has been done so far and understand how it progressed. Additionally, it offers cognitive neuroscientists a set of handles on where to start when linking neural data to cognitive models, as well as what to strive for in the long run (see also the discussion).

We then give some examples of the four categories of linking in three subfields of cognitive neuroscience from the literature. A much larger review of the literature can be viewed online as a supplement to this paper.
Finally, we will discuss the strengths and weaknesses of different points on the continuum and lay out future challenges and developments.

## Looser and tighter links

There are many approaches to linking formal models of cognition to neural data. These approaches differ in how explicit and precise the link is made between neural, physiological processes on the one hand, and cognitive, phenomenal processes on the other hand. We propose a continuum of "tightness" of linking. At the loosest level, cognitive models can be linked with neural data simply by constraining the kinds of structural assumptions allowed in the models in order to respect data about neural structures. Tighter links can be created by comparison of predictions for neural and behavioral data, or neural and behavioral model parameters. The very tightest and most explicit links are specified by "joint" models, which make quantitative predictions about both neural and behavioral data at the same time.

Table 1 Examples of four different approaches to linking models of evidence accumulation, reinforcement learning, and symbolic reasoning to neural data.provides some illustrative examples which are elaborated below. These examples highlight four commonly-used points on the continuum between loose to tight linking. Below, we first provide definitions for those four different commonly-used levels of linking. Following that, we give detailed examples of these approaches in practice, with each level of linking illustrated in up to three different research domains: perceptual decision-making; reinforcement learning, and symbolic reasoning.

- *Qualitative structural linking*: Neural data on the structure of the brain are used to constrain the structure of a cognitive model. An example of this is the leaky competing accumulator model (LCA): “the principles included in the modeling effort have neurobiological as well as computational or psychological motivation, and the specific instantiations of the principles are informed by additional neurophysiological observations” (13).
- *Qualitative predictive linking*: A cognitive model is tested using qualitative predictions about both neural and behavioral data. For example, Borst et al. (14) used the symbolic reasoning modeling framework of ACT-R to make predictions about the difference in fMRI-signals between conditions which differed in behavioral measures associated with task difficulty, separately for different brain regions: “the model does not predict a general increase in BOLD response with task difficulty; instead, it predicts lower but more persistent activation levels for the more difficult conditions in the visual and manual modules, and higher and more persistent activation levels for the more difficult conditions in the problem state and declarative memory modules” (14).
- *Quantitative predictive linking*: the predictive output of a cognitive model is quantitatively related to some aspect of neural data. In an early example of this approach, fMRI data was acquired during a Pavlovian conditioning task. The signal that was measured by fMRI was correlated with the error signal of a temporal difference (TD) algorithm performing the same task: “we used the actual output of a TD

learning algorithm to generate a PE (or δ) response at two main time points in a conditioning trial: the time of presentation of the CS and the time of presentation of the reward. The output of this algorithm was then entered into a regression model of fMRI measurements from subjects who underwent appetitive Pavlovian conditioning. This enabled us to test for brain regions that manifested a full range of TD error-related PE response'" (15).

- *Single model*: a single generative model predicts a joint distribution over both cognitive and neural data. For example, Purcell et al. (16) used single-cell recordings from monkeys: "Models using actual visual neuron activity as input predicted not only the variability in observed behavior but also the dynamics of movement neuron activity. This union of cognitive modeling and neurophysiology strengthens the interpretation of visual neuron activity as a representation of perceptual evidence of saccade target location and the interpretation of movement neuron activity as the accumulation of that evidence".

We will now present the four levels of linking in greater detail, by using examples from the fields of evidence accumulation models, value-based decision-making models, and symbolic reasoning models.

## Examples of qualitative structural linking

### Qualitative structural linking in models of evidence accumulation

Recently, attempts have been made to link evidence accumulation models with neural data. The earliest attempts, such as seminal work by Usher & McClelland (13), defined qualitative structural links. These links were structural in the sense that the constraints were applied to the structure of the model, not to the model's predictions, and the links were qualitative in the sense that the constraints revolved around the inclusion/exclusion of model elements, not to the quantitative parametric values taken. For example, the leaky competing accumulator model (LCA) of Usher & McClelland (13) specifically included structural elements such as mutual inhibition between competing accumulators. This inclusion was motivated by neural data which demonstrate the prevalence of inhibitory connections between nearby neurons within the same cortical stratum. Similarly, the LCA included passive decay of accumulated evidence, to respect the neural observation that membrane potential decays back to baseline in the absence of input. Evidence in favor of these links was inferred by the observation that the resulting cognitive model provided a good fit to behavioral data.

### Qualitative structural linking in models of reinforcement learning

The classic parallel distributed processing models provided cognitive descriptions of learning including structural constraints from neural data (PDP 17; 18). The models assumed massive parallelism and distributed information representation, reflecting key findings in the emerging neural literature on cortical structure. The models also used learning rules such as back-propagation, which were inspired by neural findings such as Hebbian plasticity.

### Qualitative structural linking in models of symbolic reasoning

The ACT-R production framework (19) is a domain-general model of human cognition. ACT-R began as a cognitive model purely aimed at behavioral data, but has since been extended in great detail to jointly consider behavioral and neural data (20; 21). The earliest linking of the ACT-R model to neural data was qualitative structural linking, which identified links between different cognitive modules in ACT-R and different brain regions. These links respected findings about the localization of brain function that were emerging at the time from the then-new method of fMRI. For example, the “visual module” of ACT-R was linked with lower occipital brain regions, and the “motor module” with motor cortices in the parietal and temporal lobes. These links defined the structure of the model and allowed the investigation of hypotheses about deficits due to brain lesions, for example.

## Examples of qualitative predictive linking

### Qualitative predictive linking in models of evidence accumulation

Hanes and Schall (22) recorded single-cell activity in the frontal eye fields (FEF) in behaving macaques. The activity of  “movement neurons” predicted the execution of saccades. Hanes and Schall showed that the ramping activity of these neurons preceding a saccade always ended with the same firing rate, but the rate of increase of firing rate was variable. The authors related these qualitative patterns to evidence accumulation models. In certain evidence accumulation models, evidence builds up gradually before a response is made, with two key properties: the rate of build-up (the “drift rate”) differs from decision to decision, but the amount of accumulated activity just before a response is issued (the “threshold”) is always the same. This is qualitatively similar to the pattern observed by Hanes and Schall.

## Qualitative predictive linking in models of reinforcement learning

The field of reinforcement learning and value-based decision-making has a long history of computational cognitive modeling (23). These computational models made it possible to design experiments that manipulated model parameters across conditions and compare the corresponding neural and behavioral data qualitatively (e.g. 24; 25). An example is given by Nieuwenhuis et al. (26),who observed that the Holroyd and Coles' model could mirror the impaired performance of older adults in a probabilistic learning task, as well as the accompanying reduced error-related negativity (ERN) measured by EEG. It could do so by varying only one parameter in the model that represents to the efficiency of dopaminergic connections to the anterior cingulate cortex (ACC).

## Qualitative predictive linking in models of symbolic reasoning

The ACT-R model assumes distinct cognitive modules that perform different parts of cognitive tasks (27). For example, the cognitive steps necessary for performing some symbolic logic operation might be modeled as involving the visual module (to perceive the stimulus), the procedural and declarative memory modules (to remember the logical rules), and the motor module (to produce the desired behavioral response). From these assumptions, ACT-R can make predictions about differences in reaction time and accuracy between conditions. Many neuroimaging studies have related cognitive ACT-R models to fMRI data to localize the cognitive modules within the brain. Such localization assumptions are linking hypotheses, and subsequent studies have used qualitative predictive approaches to test those. For example, Borst et al. (14) constructed an ACT-R model of a task where both a subtraction operation had to be performed at the same time as a text entry task. The model made priori predictions about which modules (e.g., ``Problem State'', ``Declarative Memory'', ``Manual'', and ``Visual'') would be more activated during different combinations of easy/hard versions of the two tasks. These predictions were then tested by comparing them to brain area activations during the task measured by fMRI.

## Examples of quantitative predictive linking

### Quantitative predictive linking in models of reinforcement learning

Using quantitative outputs of a computational model of cognition to predict neural activity has been a successful strategy in the study of value-based decision-making and neuroeconomics (28). Especially prominent has been the *single-trial regression approach*, in which parameters of a reinforcement learning model are estimated from choice behavior during tasks involving the learning of reward values associated with different choices. These subject-specific parameter estimates can be used to calculate estimates of the subjective values of the different choice options to the subject, for every individual trial during the experiment. These subjective, trial-by-trial values can then be used as a hypothetical cognitive signal that tracks, for example, the difference between the expected reward after a choice and the reward that was actually delivered (the so-called ``prediction error'' or ``delta'' signal). To investigate a linking hypothesis, the researcher then hypothesizes that this cognitive signal is represented in the brain, at the *bridge locus*. Neural signals from the bridge locus should correspond to the phenomelogical concept under study, and to the hypothetical cognitive signal in this case (2). For example, the bridge locus of the prediction error signal might be some area in the brain where the neural signal consistently tracks the difference between the expected reward and the actual reward in a reinforcement learning task. At a practical level, the hypothetical cognitive signal, as estimated by the reinforcement learning model, can be transformed to a hypothetical BOLD fMRI-signal, by convolution with a hemodynamic response function (29). This creates a hypothetical fMRI signal corresponding to the prediction error signal, which can be used as a regressor in a general linear model (GLM), with additional regressors for other task-related activity (for example stimulus presentation). The parameters of this GLM are then estimated for all voxels in the brain. This yields a statistical parametric mapping of the brain that shows for which areas of the brain BOLD activity correlates with the hypothetical neural signal representing stimulus value, and offers candidates for the bridge locus. This approach was used to show that the BOLD activity in orbitofrontral cortex (OFC) and in ventral striatum was correlated with the temporal difference error signal as estimated by a reinforcement learning model (15).

## Quantitative predictive linking in models of symbolic reasoning

Recent versions of the ACT-R architecture predict quantitative differences in activation in different cognitive modules during a task. These predictions can be convolved with a canonical hemodynamic response function, generating quantitative hypotheses about which areas of the brain modulate their activity in correspondence with the activity of the proposed cognitive modules in the model. Such time courses can be fitted to all the voxels throughout the entire brain. For example, Borst et al. used a multitasking paradigm, where either a subtraction or a text entry task had to be performed, while at the same time performing a listening comprehension task (30). The ACT-R model predicted, for every trial, the relative activity of ``Problem State'', ``Declarative Memory'', ``Vision'', and ``Manual'' modules. These relative activities correlated with the measured BOLD signal in corresponding brain areas.

## Quantitative predictive linking in models of evidence accumulation

Linking evidence accumulation models of speeded decision-making to neuroimaging data is more difficult than for the models of reinforcement learning and symbolic reasoning reviewed above. One reason for this is that, in order to explain random variability in reaction times, models of speeded decision-making are stochastic. Across trials, the models assume variability in the amount of evidence that is necessary to make a decision, as well as variability in the speed of evidence accumulation (as in the LBA; 31), possibly amongst even more variability (32). This means that, unlike in most reinforcement learning models, there is no one-to-one correspondence between data and the parameters of the model at the level of a single trial. This precludes the very popular single trial regression approach, at least for “out-of-the-box” evidence accumulation models, although different alternatives to resolve this issue have been proposed and performed.

One alternative is to change the unit of analysis from single trials to single subjects, focusing on the covariance of differences between subjects in neural and behavioral parameter estimates. In an fMRI study of decision-making, Forstmann et al. (33) instructed subjects to stress either the speed or accuracy of their decisions. The difference in BOLD-activity between accuracy- and speed-stressed trials in the striatum and the presupplementary motor area (pre-SMA) was correlated across subjects with the difference in model parameters related to response caution, estimated from behavioral data via the LBA model. In other words, participants who made large changes in their cognitive settings (for speed vs. caution) also showed large changes in fMRI responses, and vice versa. This provides some evidence for a role of these brain areas in setting a response threshold before a decision is made.

More recent approaches to linking evidence accumulation models to neural data start with the neural signal, and use this as input to an extended evidence accumulation model. Cavanagh et al. (34) estimated, separately for each trial in a decision-making experiment, the power in the theta frequency band from recorded EEG signals. These single-trial estimates of theta power were then used to inform parameter estimates in an extended version of the drift diffusion model (HDDM; 35). This model allowed different estimates of the threshold parameter on different trials, and a covariate model to assess the association of single-trial theta power with single-trial threshold estimates. Parameters estimated from data suggested that the coefficient of the covariate was probably larger than zero, which provides evidence that response caution (measured by the threshold parameter) is related to fluctuations in theta-power in medial prefrontal cortex.

## Single model approach

### Single model approaches in evidence accumulation

In some work in neurophysiology, the link between neural data and cognitive model is more explicit. The most complex models can take as input neural data from one source, and then predict neural data from another source, as well as behavior. Purcell et al. (16) identified and recorded from different clusters of cells in the frontal eye fields (FEF) in awake macaque monkeys during a visual search task. Some neurons in the FEF only respond to specific visual inputs (“visual” neurons), while other neurons respond only just before a saccade (“motor” neurons), and some neurons respond to both (“visuomotor” neurons). Considered from the perspective of an evidence accumulation model of decision-making, the visual neurons might be interpreted as providing a continuous, noisy, representation of decision evidence, and the motor neurons might be interpreted as the accumulators which process that evidence. Purcell et al. used the spike trains recorded from visual and visuomotor neurons as input to the accumulators of an evidence accumulation model. The model used these inputs to reliably predict the behavioral data of the monkeys (response proportions and reaction time distributions).

Purcell et al. also investigated the predictive performance of the model on neural data. For this, they used nine different architectures for evidence accumulation. These architectures differed in details like the presence or absence of leakage in the accumulation process, or mutual inhibition between accumulators. Interestingly, the response proportions and response time distributions were well explained by many of the different model architectures, even though those architectures made very different assumptions about neural structure. However, Purcell et al. showed that only one class of evidence accumulation architectures was also able to predict all the quantitative patterns in the neural data coming from the motor neurons.

### Single model approaches in symbolic reasoning

Simple symbolic reasoning models have been combined with functional neuroimaging data in a single model using Hidden Semi-Markov Models (HSMMs). Such models assume that, in order to perform a task, subjects move through a discrete set of cognitive steps, or “states”, until they finish the trial (usually by giving a response). A HSMM can be fit to both behavioral and neuroimaging data, where it is assumed that both are dependent measurements of the same sequence of states.

Anderson, Betts, Ferris, and Fincham (36; 37) first applied this approach to a dataset where students solved linear algebra problems in an MRI scanner, where every step in solving the problem was made explicit using the task interface. Given both reaction times and functional neuroimaging data, the model reliably predicted in which state of solving the linear algebra problem the subject was for a given moment in time.

## Discussion

A growing number of researchers are working towards linking formal cognitive computational models with neuroscientific data. This linking effort is made in vastly different fields of cognitive modeling, ranging from perceptual and value-based decision-making to symbolic reasoning models. These models are also linked to neural data coming from very different neuronal imaging modalities, including single-cell recordings, EEG, and fMRI. We described different kinds of linking that are applied in four discrete categories that vary along a continuum of how tight and how explicit the link between cognitive model and neural data is made. However, it is clear that even within these categories, different analyses are applied.

The great majority of studies using model-based links between neural and behavioral data, so far, are based on a regression analysis. This analysis tests for relationships between parameters estimated by a cognitive model and some aspect of a neural signal -- often a parameter estimated using a neural measurement model. In such approaches, the exact mapping from cognitive parameter to neural signal is typically left implicit, but upon closer inspection the link is almost always a linear relationship between a parameter of the cognitive model and a parameter of an easy-to-use, traditional measurement model of the neural signal. The role of the measurement model is to reduce the raw neural data to a single number (per subject, per condition, or per trial) that can be submitted to a standard regression analysis. For example, in fMRI this measurement model is most often a standard GLM used to model BOLD responses via a canonical hemodynamic response function. The GLM allows the estimation of coefficients which index the height of the hemodynamic response function, or the difference in height between conditions, and it is these coefficients that are later correlated against the parameters of a cognitive model. In EEG, the measurement model is often just the mean signal intensity in a predefined stimulus-locked time window.

The assumptions underlying these linking functions are rarely tested, even though there is ample evidence that they are probably often violated. For example, a central assumption of the canonical hemodynamic response function (HRF) model used in fMRI is that the HRF is identical across brain areas and subjects, but of course this is not true (e.g. 38). Even more problematic is that cognitive processes, as modeled by computational models, are typically assumed to modulate only the *amplitude* of the task-locked hemodynamic response. Figure 1 illustrates just how simplistic this assumption is. Contradicting this assumption, the main finding of multiple linking papers was a relationship between an estimated cognitive parameter and the *delay* (39) or *dispersion* of the HRF (30). We have also seen in our own data (e.g. 33), that cue-locked differences in the height of the HRF across conditions are often accompanied by differences in the onset-till-peak, as well as the dispersion of this HRF. These problems of violated or untested assumptions are not unique to fMRI measurement. For example, a frequent assumption in analyses of single-cell recordings is that recordings taken over different trials, cells, or conditions, are independently and identically distributed, but this is often not true (40).

The more advanced linking approaches we have reviewed, particularly the quantitative predictive approach and the single-model approach, enable future work to focus on more complex relationships and formally incorporate such links in quantitative models. In functional neuroimaging, for example, one possible strategy is to move away from voxels as the single unit-of-analysis, and move towards analyses that use anatomically-informed regions-of-interest. First of all, such a unit of analysis is often more appropriate, because anatomical boundaries are much less arbitrary than an artificial voxel-grid and they respect the close relationship between anatomical structure and function (41-43). For instance, a hypothetical relationship between degree-of-surprise and activity in the rostral part of the anterior cingulate cortex is to many neuroscientists much more interpretable than a relationship between degree-of-surprise and a blob thresholded at a z-value of 3.1, with a volume of 3300 mm$^3$ at MNI coordinate (2, 16, 32). Secondly, such an anatomically-informed approach aligns with the goal of tighter, single models that take as much neural data into account as possible. Ultimately, one would like to find bridge loci of cognitive processes that correspond to anatomical regions that have been described, validated and related to function for more than a century, rather than the smoothed and arbitrarily threshold activations blobs that frequently differ between fMRI studies (44; 45). Thirdly, there is also an important practical benefit to this: by reducing the huge dimensionality of all the voxels in the brain to the number of anatomical constructs one is interested in, the number of neural signals that need to be analyzed is reduced by multiple orders of magnitude. The corresponding reduction in computational burden allows the use of more sophisticated mathematical models, and more complete statistical treatments. Some of the more computationally demanding studies that have been published the past years would not have been feasible without dimensionality reduction of the neural data (e.g. 46-48). A possible drawback of this strategy is the lack of “negative controls”, which provide divergent validity. By exclusively focusing on a subset of a priori regions-of-interest, a researcher could potentially miss other brain areas that should be of interest. Clearly, the set of brain areas that is investigated should be narrowed down only after more exploratory research uncovered the set of potential bridge loci within all brain areas.

Even if these quantitative links between model parameters and neural signals are made more explicit, researchers must still remain circumspect in the conclusions they draw from any of these statistical links they find between the two. Clearly, computational models that were developed in cognitive psychology are there to explain *cognition*. At best, they give a formalization of the kind of cognitive processes the human brain can perform, how they differ in different circumstances, and formalize differences in cognition across subjects. They describe the algorithms that take place in the brain and which quantities it therefore has to compute, but these models usually remain agnostic about the precise implementation of these algorithms at the level of neural signals (3; 7; 49; 50).

When a brain region is identified that shows a correlation between neural signal and a cognitive parameter, this area may be involved in computing the quantity that corresponds to that parameter. Nevertheless, it is still very unlikely that such a cognitive parameter is a cognitive process in itself, and that is has a simple one-to-one, cognitive-process-to-brain-area mapping (50; 51). This is because many other hypotheses can be generated which are consistent with the observed link, but differ in the mechanics that explain the link (e.g., perhaps the identified region simply relays or mirrors the signal of interest). Finding relationships between cognitive models and neural measurements is just a first step toward more detailed neurocomputational models (e.g. 8).

## When is a linking approach not good enough?

It is difficult to avoid the appearance of value judgments in our proposed continuum of loose-to-tight linking. However, we would like to stress that it is not true that tighter linking is always preferable to looser linking. The ideal point on the continuum depends on many things, but most importantly on the status of the formal quantitative models in the research paradigm of interest. Tighter linking approaches are only feasible when there exist well-understood and settled quantitative models of both the behavioral data and the neural data in the paradigm of interest. Even more, these models must be computationally or analytically tractable, for any level of linking, and they should preferably specified at the level of distributions and likelihood functions for the tightest possible linking. Where these assumptions are not met, linking approaches beyond the simplest qualitative structural or qualitative predictive are not likely to succeed (e.g., this was the case for models of perceptual decision-making in the late 1990s and early 2000s). Conversely, in areas for which tractable quantitative models do exist, it is incumbent on researchers to strive for the tightest possible linking approaches (e.g., it is no longer reasonable for perceptual decision-making models to be linked to neural data using qualitative approaches).

We propose the following guidelines for researchers searching for the right way to link their behavioral and neural data via cognitive models. First of all, when a researcher is interested in how a cognitive model relates to computations performed in the brain, it is always a good idea to come up with qualitative predictions. One should always ask themselves the question of what aspects of a model correspond to structural aspects of the brain (qualitative structural linking), as well as how variability in the parameters of the model across subjects, conditions or trials, would relate to differences in neural functioning (qualitative functional linking). These intuitions can then be tested empirically by clever experimental design, modulating some aspect of either the neural or cognitive domain and measuring the corresponding change in the other one.

Only after such qualitative predictions have been made and possibly tested, one should wonder if these predictions can be phrased in a quantitative way. Is there a neural measure that possibly corresponds to variability in the neural process we hypothesize? And can we come up with a parameter estimate from the model that it can be related to? If so, a wealth of literature has shown relatively straightforward methods of relating the two to each other, which we have reviewed under quantitative functional linking. If it’s possible to use more elaborate models of the neural data, this is usually a good idea, especially in an explorative setting, because it allows for relaxing and testing the assumptions about the links between neural and cognitive models (e.g., maybe the dispersion of the HRF in the striatum and not so much the height is modulated by response caution).

Third, only when such exploratory, but quantitative, efforts have been successful, one can start thinking about how to unify the findings into a single model. This is a very challenging enterprise, as we understand little about how the brain actually computes anything, but recent successes suggest that concepts from cognitive and neural models can to a surprising extent just be equated. For example, the decision-related signal in single neurons in LIP is not just a correlate of the decision variable, but may be the decision variable itself (16). And maybe cognitive states and neural states during problem solving are two sides of the same coin (37). While the exercise of developing single models explaining both behavioral and neural data has only just begun, it will be very interesting to see what new models the scientific community can develop and extend in the coming future.

## Conclusions

Linking formal cognitive models to neural data can improve our understanding of how the brain functions. However, the precise technical details and philosophical approach to tackling this linking problem can vary greatly. We have focused on one important attribute of linking: the degree of tight, quantitative, and explicit hypothesizing. In each field of cognitive research, the earliest approaches to linking behavioral and neural data are typically qualitative. As knowledge is accumulated, and as formal models become more settled, linking approaches become more quantitative and more explicit. Even in that case, however, model parameters are most frequently estimated separately for the behavioral and neural models, and the two models are brought together only at the very end, in a simple linear regression on model parameters. Future work should explicitly model how differences in cognitive parameters modulate differences in the signal in the neural domain, thereby acknowledging the richness of the data in the neural domain and exploit more than the most common parameter of the most simple measurement model in that domain.

## Financial Disclosures

Dr. Birte U. Forstmann, Dr. Scott Brown and Gilles de Hollander reported

no biomedical financial interests or potential conflicts of interest.

1. Forstmann BU, Wagenmakers E-J, Eichele T, Brown S, Serences JT (2011): Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends in Cognitive Sciences*. 15: 272–279.
2. Teller DY (1984): Linking propositions. *Vision research*. 24: 1233–1246.
3. Schall JD (2004): On Building a Bridge Between Brain and Behavior. *Annu Rev Psychol*. 55: 23–50.
4. Lewandowsky S, Farrell S (2010): *Computational modeling in cognition: Principles and practice*. Sage.
5. Forstmann BU, Wagenmakers E-J (2015): *An Introduction to Model-Based Cognitive Neuroscience*. (B. U. Forstmann & E.-J. Wagenmakers, editors). Springer, pp 1–358.
6. Hawkins GE, Boekel W, Heathcote A, Forstmann BU (2015): Toward a model-based cognitive neuroscience
of mind wandering.
. *Neuroscience*. 1–39.
7. Marr D (1982): *Vision: A computational approach*. Freeman & Co., San Francisco.
8. Bogacz R, Wagenmakers E-J, Forstmann BU, Nieuwenhuis S (2010): The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences*.

33: 10–16.

9. Robinson DA (1992): Implications of neural networks for how we think about brain function. *Behavioral and brain sciences*.
10. Steingroever H, Wetzels R, Wagenmakers E-J (n.d.): Bayes factors for reinforcement-learning models of the Iowa gambling task. *Decision*.
11. Turner BM, Sederberg PB (2014): A generalized, likelihood-free method for posterior estimation. *Psychon Bull Rev*. 21: 227–250.
12. Hawkins GE, Forstmann BU, Wagenmakers E-J, Ratcliff R, Brown SD (2015): Revisiting the Evidence for Collapsing Boundaries and Urgency Signals in Perceptual Decision-Making. *Journal of Neuroscience*. 35: 2476–2484.
13. Usher M, McClelland JL (2001): The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*. 108: 550–592.
14. Borst JP, Taatgen NA, Stocco A, van Rijn H (2010): The Neural Correlates of Problem States: Testing fMRI Predictions of a Computational Model of Multitasking. (B. J. Harrison, editor) *PLoS ONE*. 5: e12966.
15. O'Doherty JP, Dayan P, Friston KJ, Critchley H, Dolan RJ (2003): Temporal difference models and reward-related learning in the human brain. *Neuron*. 38: 329–337.
16. Purcell BA, Heitz RP, Cohen JY, Schall JD, Logan GD, Palmeri TJ (2010): Neurally constrained modeling of perceptual decision making. *Psychological Review*. 117: 1113–1143.
17. Rumelhart DE, Hinton GE, McClelland JL (1986): A General Framework for Parallel Distributed Processing. In: Rumelhart DE, McClelland JL, the PDP Research Group, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol 1)*. Cambridge, MA: MIT Press, pp 45–76.
18. Rumelhart DE, Hinton GE, Williams RJ (1986): Learning Internal Representations by Error Propagation. In: Rumelhart DE, McClelland JL, the PDP Research Group, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol 1)*. Cambridge, MA: MIT Press, pp 318–362.
19. Anderson JR (1992): Automaticity and the ACT* Theory. *American Journal of Psychology*. 105.
20. Sohn MH, Goode A, Stenger VA, Carter CS, Anderson JR (2003): Competition and representation during memory retrieval: Roles of the prefrontal cortex and the posterior parietal cortex. *Proceedings of the National Academy of Sciences*. 100: 7412–7417.
21. Qin Y, Sohn MH, Anderson JR, Stenger VA, Fissell K, Goode A, Carter CS (2003): Predicting the practice effects on the blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task. *Proceedings of the National Academy of Sciences*. 100: 4951–4956.
22. Hanes DP, Schall JD (1996): Neural control of voluntary movement initiation. *Science*. 274: 427–430.
23. Sutton RS, Barto AG (1998): *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press.
24. Berns GS, McClure SM, Pagnoni G, Montague PR (2001): Predictability modulates human brain response to reward. *Journal of Neuroscience*. 21:

2793–2798.

25. Knutson B, Adams CM, Fong GW, Hommer D (2001): Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*. 21: RC159.
26. Nieuwenhuis S, Ridderinkhof KR, Talsma D, Coles MGH, Holroyd CB, Kok A, van der Molen MW (2002): A computational account of altered error processing in older age: dopamine and the error-related negativity. *Cognitive, Affective, & Behavioral Neuroscience*. 2: 19–36.
27. Anderson JR (2007): *How can the human mind occur in the physical universe?* New York, NY, USA: Oxford University Press.
28. Corrado GS, Sugrue LP, Brown JR, Newsome WT (2009): The Trouble with Choice: Studying DecisionVariables in the Brain. In:. *Neuroeconomics*. pp 1–19.
29. Glover GH (1999): Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*. 9: 416–429.
30. Borst JP, Taatgen NA, van Rijn H (2011): Using a symbolic process model as input for model-based fMRI analysis: Locating the neural correlates of problem state replacements. *NeuroImage*. 58: 137–147.
31. Brown SD, Heathcote A (2008): The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*. 57: 153–178.
32. Ratcliff R, McKoon G (2008): The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*. 20: 873–922.
33. Forstmann BU, Dutilh G, Brown S, Neumann J, Cramon Von DY, Ridderinkhof KR, Wagenmakers E-J (2008): Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*. 105: 17538–17542.
34. Cavanagh JF, Wiecki TV, Cohen MX, Figueroa CM, Samanta J, Sherman SJ, Frank MJ (2011): Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Publishing Group*. 14: 1462–1467.
35. Wiecki TV, Frank MJ (2013): A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*. 120: 329–355.
36. Anderson JR, Betts S, Ferris JL, Fincham JM (2010): Neural imaging to track mental states while using an intelligent tutoring system. *Proc Natl Acad Sci USA*. 107: 7018–7023.
37. Anderson JR (2012): Tracking problem solving by multivariate pattern analysis and Hidden Markov Model algorithms. *Neuropsychologia*. 50: 487–498.
38. Gonzalez-Castillo J, Saad ZS, Handwerker DA, Inati SJ, Brenowitz N, Bandettini PA (2012): Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proceedings of the National Academy of Sciences*. 109: 5487–5492.
39. Ho TC, Brown S, Serences JT (2009): Domain General Mechanisms of Perceptual Decision Making in Human Cortex. *Journal of Neuroscience*. 29: 8675–8687.
40. Aarts E, Verhage M, Veenvliet JV, Dolan CV, van der Sluis S (2014): A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Publishing Group*. 17: 491–496.
41. Alkemade A, Schnitzler A, Forstmann BU (2013): Anatomy and function of

the human subthalamic nucleus. 1–22.
42. Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, *et al.* (2009): Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*. 106: 13040–13045.
43. Lohmann G, Stelzer J, Neumann J, Ay N, Turner R (2013): “More Is Different” in Functional Magnetic Resonance Imaging: A Review of Recent Data Analysis Techniques. *Brain Connectivity*. 3: 223–239.
44. Derrfuss J, Mar RA (2009): Lost in localization: The need for a universal coordinate database. *NeuroImage*. 48: 1–7.
45. Brodmann K (1909): Vergleichende Lokalisationslehre der Groshirnrinde. *Leipzig: Barth*.
46. Turner BM, van Maanen L, Forstmann BU (2015): Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychological Review*. 122: 312–336.
47. Wiecki TV, Sofer I, Frank MJ (2013): HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform*. 7: 14.
48. Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D (2015): fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning. *Journal of Neuroscience*. 35: 485–494.
49. Moore EF (1956): Gedanken-experiments on sequential machines. *Automata studies*. 34: 129–153.
50. Mars RB, Shea NJ, Kolling N, Rushworth MFS (2012): Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *The Quarterly Journal of Experimental Psychology*. 65: 252–267.
51. O’Reilly JX, Mars RB (2011): Computational neuroimaging: localising Greek letters? Comment on Forstmann et al. *Trends in Cognitive Sciences*. 15: 450.
52. Frank MJ, Claus ED (2006): Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*. 113: 300–326.
53. van Maanen L, Brown SD, Eichele T, Wagenmakers E-J, Ho T, Serences J, Forstmann BU (2011): Neural Correlates of Trial-to-Trial Fluctuations in Response Caution. *Journal of Neuroscience*. 31: 17488–17495.
54. Purcell BA, Schall JD, Logan GD, Palmeri TJ (2012): From salience to saccades: multiple-alternative gated stochastic accumulator model of visual search. *Journal of Neuroscience*. 32: 3433–3446.

# Figures

Figure 1 Standard canonical HRF model of BOLD-activity. In regression-based approaches, it is assumed that the only way the BOLD-response is modulated is that the canonical shape is multiplied by some factor β.