Trends in Cognitive Sciences

CellPress

Opinion

The Importance of Falsification in Computational Cognitive Modeling

Stefano Palminteri,[1,2,*,3] Valentin Wyart,[1,2,*,3] and Etienne Koechlin[1,2,*]

In the past decade the field of cognitive sciences has seen an exponential growth in the number of computational modeling studies. Previous work has indicated why and how candidate models of cognition should be compared by trading off their ability to predict the observed data as a function of their complexity. However, the importance of falsifying candidate models in light of the observed data has been largely underestimated, leading to important drawbacks and unjustified conclusions. We argue here that the simulation of candidate models is necessary to falsify models and therefore support the specific claims about cognitive function made by the vast majority of model-based studies. We propose practical guidelines for future research that combine model comparison and falsification.

Trends

Computational modeling has grown exponentially in cognitive sciences in the past decade.

Model selection most often relies on evaluating the ability of candidate models to predict the observed data.

The ability of a candidate model to generate a behavioral effect of interest is rarely assessed, but can be used as an absolute falsification criterion.

Recommended guidelines for model selection should combine the evaluation of both the predictive and generative performance of candidate models.

Complementary Roles of Comparison and Falsification in Model Selection

Computational modeling has grown considerably in cognitive sciences in the past decade (Figure 1A). Computational models of cognition are also becoming increasingly central in neuroimaging and psychiatry as powerful tools for understanding normal and pathological brain function [1–5]. The importance of computational models in cognitive sciences and neurosciences is not surprising; because the core function of the brain is to process information to guide adaptive behavior, it is particularly useful to formulate cognitive theories in computational terms [6,7] (Box 1). Similarly to cognitive theories, computational models should be submitted to a selection process. We argue here that the current practice for model selection often omits a crucial step: **model falsification** (see Glossary).

One universally recognized heuristic for theory selection is Occam's law of parsimony: '*pluralitas non est ponenda sine necessitate*' (plurality is never to be posited without necessity). This principle dictates that among 'equally good' explanations of data, the less complex explanation should be held as true. More formally, a trade-off exists between the complexity of a given model (which specifically grows with its number of 'free' and adjustable parameters) and its goodness-of-fit (the likelihood of the observed data given the model). Different quantitative criteria (e.g., the Bayesian information criterion, Bayes factor, and other approximations of the model evidence) have been proposed to take **model parsimony** into account when comparing different models. These criteria are based on the **predictive performance** of a model, in other words its ability to predict the observed data [8–11]. We refer to them as relative comparison criteria because they imply no absolute criterion for model selection or rejection. Following these criteria, the 'winning' (or 'best') model is the model with the strongest evidence (i.e., trading off goodness-of-fit with complexity) compared to rival models [8,12]. Various statistical methods can then be used to test whether there is significantly stronger evidence in favor of the winning model than rival models.

[1]Laboratoire de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, Paris, France
[2]Institut d'Étude de la Cognition, Departement d'Études Cognitives, École Normale Supérieure, Paris, France

*Correspondence:
stefano.palminteri@ens.fr (S. Palminteri),
valentin.wyart@ens.fr (V. Wyart),
etienne.koechlin@ens.fr (E. Koechlin).

Box 1. Delineating Computational Modeling Approaches

In cognitive sciences, computational models can be used either as analytical tools for analyzing empirical data or as instantiations of cognitive hypotheses. In the first case, the typical results consist of comparing model parameters across conditions or subjects [27], in other words computational models are treated as statistical models, similar to multiple regressions. In this approach, model comparison is not crucial because the models are not instantiations of cognitive theories.

As instantiations of cognitive theories, computational models can target different levels of description. Clearly identifying the target level should precede a model comparison analysis. A key distinction is between aggregate versus mechanistic models [9]. Aggregate models aims to describe average behaviors using a synthetic mathematical model, such as an exponential learning curve [28]. Mechanistic models aim to explain how behaviors are generated, such as the 'delta rule' in reinforcement learning [29]. Because these two types of models do not target the same level of description, there is no reason to arbitrate between aggregate versus mechanistic models. For example, an aggregate exponential learning curve could be derived formally from a 'delta rule' such that both models are equivalent. The distinction between aggregate and mechanistic models has been further developed by Marr [6], who proposed three distinct levels of description. The 'computational' level corresponds to the goal of the model. The 'representational' or 'algorithmic' level corresponds to a computational model formulated in terms of the mathematical operations (algorithms) that transform inputs into outputs (representations). Finally, the 'physical' or 'implementational' level corresponds to the biological implementation of a computational model in the brain (or an artificial device). Again, there is no reason to arbitrate between models across levels of description. In addition, the comparison of models has different meanings at the 'computational', 'algorithmic', and 'physical' levels. At the 'computational' level, model comparison informs about the actual task that subjects realize, whereas at the 'algorithmic' level model comparison informs about the way subjects perform this task [30]. Because simulating a model requires an algorithm to be specified, it is essential to clearly mention whether the model reflects a hypothesis at the computational or algorithmic level.

Glossary

Generative performance: the ability of a given model to generate the data. The generative performance is evaluated by comparing model simulations to the actual data. For this comparison both frequentist and Bayesian statistics can be used.

Model falsification: showing through model simulations that a given model is not able to generate a specific behavioral effect of interest. The simulated data should be generated using the best-fitting parameter values. Ideally, this 'model falsification' step should include two related results: (i) the behavioral phenomenon is not detectable in the simulated data, and (ii) a significant difference between observed and simulated data should be detected. Statistical tests used in model falsification could belong to both Bayesian and frequentist statistical traditions.

Model generalizability: evaluating the ability of the best-fitting model and the best-fitting parameters to predict the data out-of-sample.

Model parsimony: the opposite of model complexity, which is classically indexed by the number of free/adjustable parameters of a given model.

Model recovery: a procedure consisting of generating synthetic data from a known candidate model and subsequently verifying the ability of a relative model comparison criterion to correctly identify the model used to generate the synthetic data.

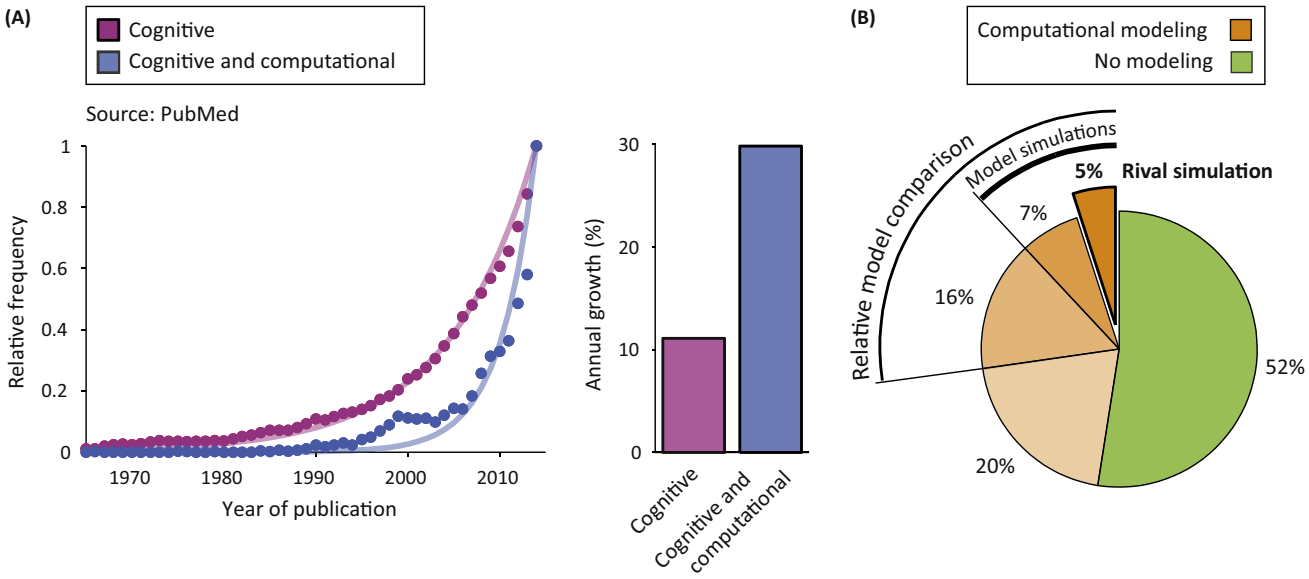
Predictive performance: the ability of a given model to predict the data. Typically the predictive performance is instantiated by the likelihood of observing the experimental data given the model. The predictive performance of models is used to calculate various approximations of the model evidence (e.g., BIC, AIC, and others).

However, contemporary epistemology recognizes that parsimony is not the heuristic required for selecting theories. Proposing a new theory requires researchers to report experimental data that contradict (or 'falsify') an existing theory, whereas the new theory is able to account for these data (along with previous ones) [13,14]. Falsifying a cognitive model relies on showing that it is unable to account for a specific behavioral (or neural) effect of interest. We propose to define the inability to account for a specific effect of interest as an absolute rejection criterion during model selection [15]. The ability of a cognitive model to reproduce (or not) the effect of interest – which we refer to as its **generative performance** – needs to be assessed by simulating the model and comparing the simulated data to the observed data. Various statistical approaches – both frequentist (e.g., *t*-tests, analyses of variance) and Bayesian – can then be used to test whether the simulated and observed effects are different, in which case the simulated model can be rejected outright irrespective of its comparison to other models.

Relative comparison criteria are inappropriate for falsifying models because (i) they focus on relative evidence in favor of the winning model and against rival models, and (ii) they are blind to the ability of candidate models to produce any specific effect of interest found in the data.

To illustrate the complementary roles of model comparison (based on model fitting) and model falsification (based on model simulations), we sketch two recent examples taken from the learning and decision-making literature [16,17].

In the first study, the authors studied the origin of human choice variability in a canonical decision-making task involving the categorization of sequences of visual stimuli of variable lengths (Figure 2A) [16]. They compared a standard model in which variability arose from a noisy response selection process to a new model in which variability arose from errors in the inference process. In this example, the two models had the same complexity – in other words one variability parameter located either at the inference or response selection stages. The authors first assessed the predictive performance of the two models, which provided substantial evidence in favor of the 'noisy inference' model. Then, to determine why the 'noisy inference'



Trends in Cognitive Sciences

Figure 1. The Exponential Increase of Computational Model-Based Cognitive Neuroscience and Current Practice in a Representative Sample. (A) The curves on the left show the relative frequency of PubMed entries for 'cognitive' (in red) and 'cognitive and computational' (in blue) as a function of the year. Their frequencies are calculated relative to the number of entries of 2014, which are therefore normalized to 1 for both curves. The bars on the left represent the estimated annual growth of the best-fitting exponential curve. (B) Survey of recent literature in the authors' database. 'Computational modeling' denotes studies reporting a computational model-based analysis. 'Relative model comparison' denotes studies reporting a model selection step that implies relative model comparison criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) or a similar method. 'Model simulations' are studies reporting, at least, the model simulation of the 'winning' model according to relative model comparison. 'Rival simulation' includes studies reporting the simulation of the 'winning' and rival model(s). Note that the presence of rival model simulation represents no guarantee that any statistical analysis is subsequently performed to quantitatively assess the 'similarity' of the model simulations to the actual data. Screened studies (n ~ 300) were published since 2009 in *Nature*, *Science*, *Nat. Neurosci.*, *Neuron*, *Proc. Natl. Acad. Sci. USA*, *PLOS Biol.*, and *J. Neurosci*.

model outperformed the 'noisy selection' model, they simulated the predictions of the two models in terms of choice variability as a function of sequence length. Crucially, the simulations of the two models diverged qualitatively on this behavioral dimension (Figure 2B): the 'noisy selection' model predicted a constant choice variability, whereas the 'noisy inference' model predicted a linear increase of choice variability as a function of sequence length. The human data showed the same linear increase of choice variability as the 'noisy inference' model, and differed statistically from the simulations of the 'noisy selection' model.

In the second study, the authors studied whether humans learn subjective values in an absolute or a context-dependent scale during reinforcement learning (i.e., learning by trial-and-error) [17] (Figure 2C). To do so they devised two experimental conditions corresponding to different learning contexts: one in which choice options were associated with monetary gains, and the other with monetary losses. They compared a standard reinforcement learning model, in which subjective values are learned on an absolute scale, to a new model in which values are learned in a context-dependent manner. In this example, note that the 'contextual' model included an additional 'context value' parameter compared to the 'absolute' model. As in the first study, the authors first assessed the predictive performance of the two models, which provided substantial evidence in favor of the 'contextual' model. Crucially, the simulations of the two models diverged significantly in terms of subjective preference ratings measured after the learning task (Figure 2D): while the 'absolute' model predicted subjective values to grow monotonically with objective values, the 'contextual' model predicted a context-dependent value distortion. The

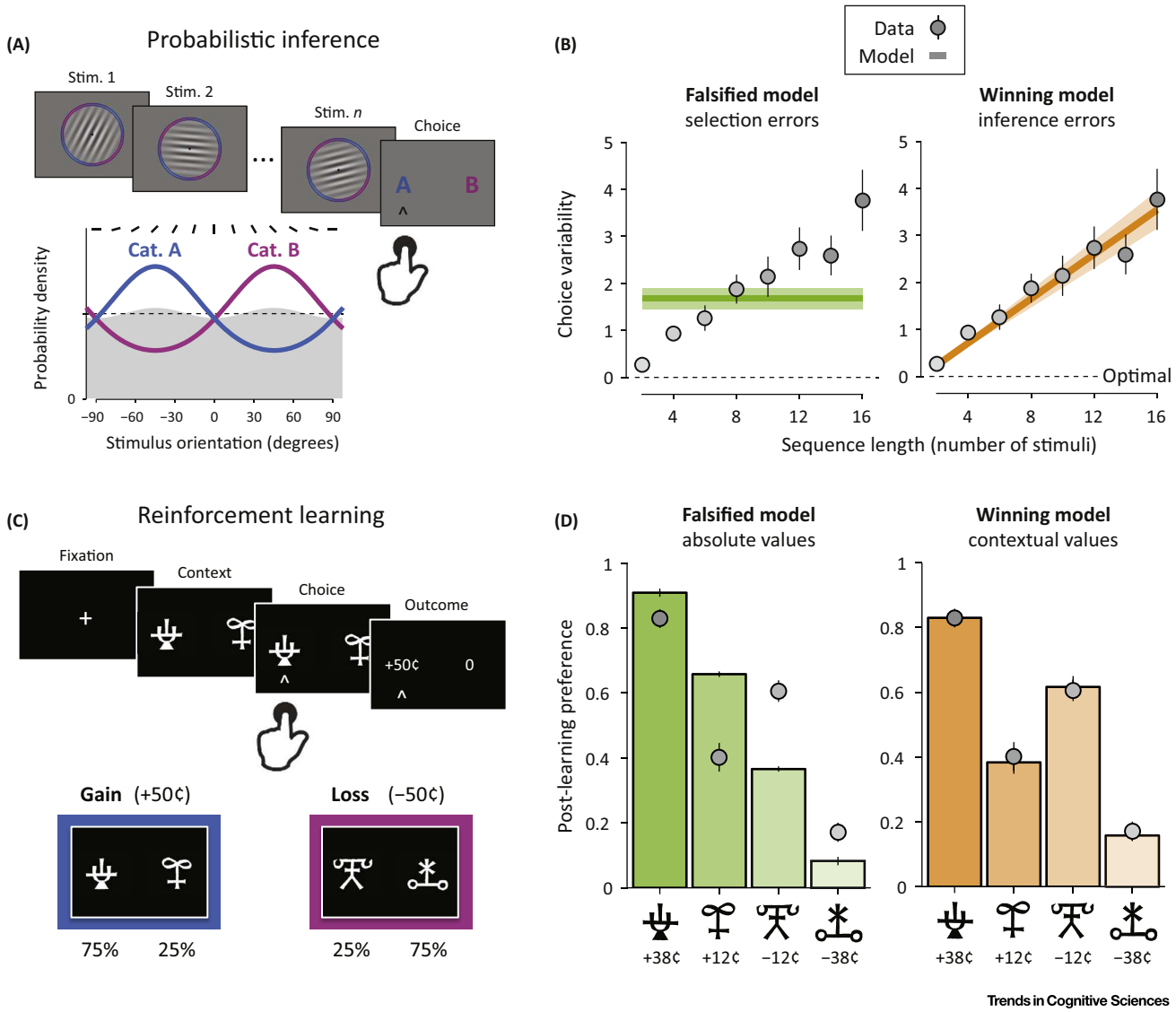


Figure 2. Concrete Examples of Model Falsification. (A) Probabilistic inference task: each trial consisted of a sequence of 2–16 stimuli (stim.) drawn from a generative probability distribution centered on one among two cardinal orientations. (B) Observed (grey dots) and model simulated (colored lines) choice variability in the probabilistic inference task as a function of the sequence length. (C) Reinforcement learning task: each trial consisted of two stimuli associated with different probabilities of winning and losing money. (D) Observed (grey dots) and model simulated (colored bars) post-learning preference as a function of the stimulus value.

human data showed the same context-dependent value distortion as the 'contextual' model, and differed statistically from the simulations of the 'absolute' model.

These two studies illustrate the different levels of interpretation that can be reached from relative model comparison alone and from model falsification through model simulations. In contrast to relative criteria, the comparison of model simulations to the observed data can lead to the outright rejection of a candidate model independently of the capacity of any other candidate model to account for a behavioral effect of interest (i.e., an absolute rejection criterion). In both studies the authors presented two complementary findings: (i) a standard model is falsified by its inability to reproduce a behavioral effect of interest, and (ii) a new model is proposed based

on its ability to reproduce the same effect. Importantly, note that only the first result is definitive. Indeed, while future research may potentially reveal other behavioral effects that are not explained by the 'noisy inference' and the 'contextual' models, the falsification of the 'noisy selection' and the 'absolute' model will still hold.

Model Selection in Cognitive Sciences: Current Practice

Although the use of relative model comparison criteria for selecting models is becoming a standard practice in computational modeling studies of cognitive functions (e.g., perception, learning, decision-making), the simulation of candidate models is rarely performed to support claims about model selection. A survey of several studies recently published in six high-impact journals in the fields of learning and decision-making (Figure 1B) illustrates the relative lack of model simulations in the literature. In this sample, most studies undertake computational cognitive modeling. Among these studies, more than 50% use a relative model comparison criterion for selecting models. However, fewer than 20% of the remaining studies simulate models. We showed above that not comparing model simulations is problematic when the cognitive model is supposed to account for a specific behavioral phenomenon. This issue is particularly relevant given that (i) parameter and **model recovery**, as well as out-of-sample likelihood estimation, are often both omitted, and (ii) without these quality checks one can never be certain that relative model comparison is not biased in favor of the winning model [18].

Our argument predominantly stands for typical modeling studies that proceed as follows. First, a task is designed to elicit a detectable behavioral effect of interest for arbitrating between two (or more) competing hypotheses about a cognitive process. These hypotheses are formulated in terms of distinct computational models. The eventual aim is to decide which of two hypothetical models accounts for the behavioral effect of interest. Thus, failing to reproduce the effect of interest represents the model rejection criterion that falsifies one hypothesis.

Second, one usually determines the model free parameters that maximize the likelihood of the data given the model (referred to as model fitting). The likelihood is then used to calculate a relative quality-of-fit criterion (e.g., the Bayesian information criterion) for comparing candidate models and ultimately identifying a 'winning' model. Most studies then omit the simulation of candidate models, implicitly assuming that this relative model comparison procedure is sufficient to conclude that the winning model provides a better account of the behavioral effect of interest. This conclusion is not necessarily warranted, however, because the ability to predict and the ability to generate an effect of interest are not necessary related (simulations are presented in Box 2).

However, very few model-based studies perform model simulation. These studies overlook the crucial complementarity between model parsimony and falsification. Whereas relative model comparison enables the identification of the most likely model among tested candidates, only model simulation can both provide the causes of the (good or bad) quality-of-fit and inform the relationship between the model and a behavioral effect of interest. Provided that this relationship is not established by analyzing the model simulations, there are no reasons to accept a model as an account of this empirical phenomenon. In conclusion, the frequent absence of model simulations in empirical studies involving computational models (up to 82% in our survey) leads to a situation in which a sizable fraction of studies incorrectly reject or accept particular computational theories with no compelling evidence.

When dealing with model comparison and selection, an important issue is to define the model space, in other words the set of candidate computational models to be compared in a given study. There is of course no theoretical upper limit to the model space size – even in simple paradigms. In line with the arguments presented above, increasing the model space size does

not make relative model comparison stronger. However, the arguments presented above allow us to propose at least a lower limit: the model space must contain (i) at least one model that is able to, and (ii) one model that cannot produce the effect of interest, when assessed through model simulation. Finally, the logic of scientific progress is to replace old theories with new ones as soon as old theories fail to account for new experimental findings [13]. Accordingly, the model space may include one 'reference' model that corresponds to a commonly accepted hypothesis together with at least one 'target' model that corresponds to an alternative hypothesis and which reproduces the behavioral effect of interest.

Model Selection in Cognitive Sciences: Proposed Guidelines

In this section we propose some basic guidelines for model selection in cognitive science that consider both model parsimony and model falsification. These guidelines combine both relative model comparison criteria and model simulations (see [16,19,20] as examples of studies that include all these steps).

(i) Given a cognitive process of interest, define a task that is intended to challenge different computational models that describe this process. Specifically, the protocol should be built to reveal at least one behavioral effect of interest, allowing discrimination between the different hypothetical models.

(ii) Simulate *ex ante* the two (or more) competing computational models across a large range of parameters to ensure that the task allows discrimination between the models: the models should exhibit distinct behaviors along the cognitive dimension(s) of interest over (preferentially large) parameter ranges. Concomitantly verify that the relative model comparison criteria allow correct recovery of each true generative model of these various simulated behaviors as the 'winning model' (a procedure coined as model or parameter recovery). Alter the task and return to step 1 or amend the model space as long as the candidate models fail to pass step 2.

(iii) Run the experiment and collect the data.

(iv) Verify the presence of behavioral effect(s) of interest in the data.

(v) Fit the competing candidate models to the data both to obtain the best-fitting model parameter values and to identify the most parsimonious model using relative model comparison criteria. The preferable approach consists in estimating the model likelihood

Box 2. Model Predictions and Simulations Are Not Necessarily Related

Relative model comparison criteria evaluate the 'predictive performance' of each candidate model, in other words the likelihood of observing the data given each model. However, a candidate model can exhibit high predictive performance but fail to reproduce a behavioral effect of interest through model simulations, in other words it displays poor 'generative performance'. This discrepancy is well captured by an analogy to weather forecasts [9]. A reasonable predictive model follows the hypothesis that 'tomorrow's weather will be similar to today's'. Given the large temporal autocorrelation of weather conditions, this simplistic model predicts weather above chance. Nonetheless, this model is by no means informative about the mechanisms governing the weather and is thus unable to simulate its evolution in the absence of past observations. Unfortunately, such temporal autocorrelation is often present in human behavior, especially across sequential decisions (previously rewarded choices tend to be repeated even when this strategy is suboptimal [31]). To illustrate this point, we generated synthetic choice data from a canonical 'reversal learning' task in which subjects must establish which of two symbols is rewarded more frequently when chosen (with probability $P_{rew} = 0.8$, the other being rewarded with probability $P_{rew} = 0.2$). Unknown to the subjects, the more frequently rewarded action reverses after 15 trials, thereby inciting subjects to adapt their choice behavior. The simulated choice data (black lines) was obtained from a standard 'Q-learning' model [32] with a learning rate of 0.1 and whose latent subjective values are corrupted by additive Gaussian noise with a standard deviation of 0.2. Based on this canonical task and synthetic choice data, we compared the predictive and generative performance of two assumedly simplistic candidate models of choice behavior (Figure IA): (i) a win–stay/lose–switch (WS-LS) model (orange lines) which chooses on the basis of the previous outcome: $action_1$ with probability P if $action_1$ was rewarded on the previous trial or $action_2$ was not rewarded on the previous trial, and $action_2$ with probability $1 - P$ otherwise; (ii) a repetition model (green lines) similar to the simplistic weather forecast model, which chooses solely on the basis of its previous choice: $action_1$ with probability P if $action_1$ was chosen on the previous trial, and $action_2$ with probability $1 - P$ otherwise. Fitting the simulated choice data with the two candidate models through maximum likelihood estimation (MLE) revealed that the repetition model explained the data better than

the WS-LS model, even though the repetition model is entirely blind to reward history and chooses only on the basis of its own previous action (Figure IA, left panel). By contrast, simulating both models *de novo* using their best-fitting parameters exposed the complete inability of the repetition model to track reward history and reversal (Figure IA, right panel). This simple example illustrates the extent to which the predictive performance of a model can be dissociated from its generative performance, and can lead to erroneous conclusions concerning the ability of a model to capture a behavioral effect of interest.

To further highlight the importance of identifying a behavioral effect of interest for model selection, we consider the performance of two more sophisticated models in our previous example (Figure IB). In this case, the simulated date were fitted with a standard Q-learning model (orange lines) and a more sophisticated version of the repetition model, the 'response history' model, which bases current choices on a weighted average of recent choices. Whereas the Q-learning model fits accurately the observed data, both before and after reversal, the repetition history model fails to do so. However, and surprisingly, relative model comparison indicates that the response history model outperforms the Q-learning model at predicting the observed data. This discrepancy can be understood by plotting another behavioral effect – the mutual dependency across successive choices [20]. The observed data show an elevated mutual dependency as a result of internal corruptive noise, which is not captured at all by the standard Q-learning model.

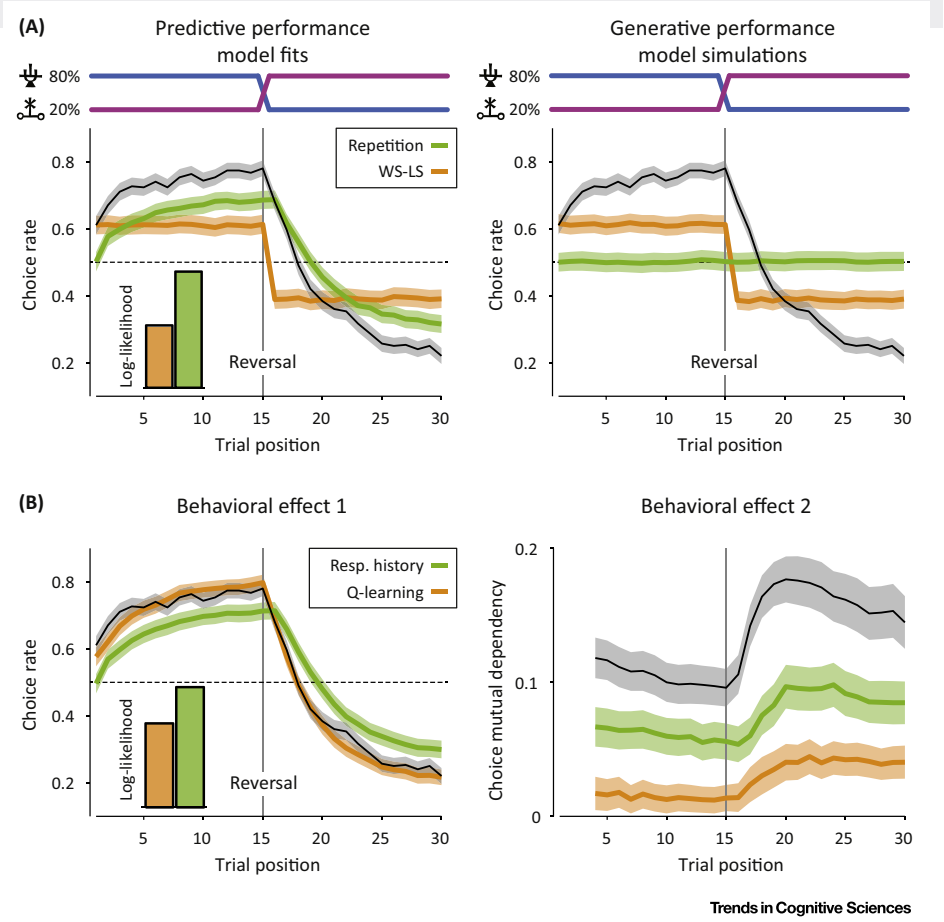


Figure I. Model Predictions versus Model Simulations. (A) Predictive versus generative performance. The black curves represent the data (noisy Q-learning). The colored curves represent the model fits (leftmost panel) and the model simulations (rightmost panel) of a repetition (green) model and a WS-LS (orange) model. (B) Predictive performance as a function of different behavioral effects of interest. The black curves represent the data (noisy Q-learning). The colored curves represent the model fits of the choice rate (leftmost panel) and the choice mutual dependency (rightmost panel) of a response history model (green) model and a simple Q-learning (orange) model. In (A) and (B) the inset represents the result of the relative mode comparison: the winning model has the higher log-likelihood.

'out-of-sample' because it does not rely on any particular approximation of the model evidence (i.e., **model generalizability**) [18].

(vi) Simulate the models *ex post* with their best-fitting parameter values to verify that, for the retrieved set of model free parameters [21], only the behavior of the best-fitting model, and ideally not the rival one(s), can reproduce the behavioral effect of interest.

In practice, computational model-based studies in cognitive sciences fall into two categories. The first category, 'data first' [17,22], includes the studies that seek a computational explanation for a previously documented empirical phenomenon of interest. The second category, 'model first' [23], includes the studies that conduct experiments to discriminate between competing computational hypotheses. The 'data-first' category can only involve steps (iv) to (vi), given that the data precede computational modeling. In the 'model first' cases, the computational model itself can provide the key insight to identify the behavioral phenomena of interest by, for example, aligning the data to events that are not cued by the task, but are instead predicted by the model [24].

For 'model-first' studies, the identification of a falsifiable model prediction can be difficult owing to multiple interactions among different parameters or small effect sizes on average behavior. In such cases, the capacity to recover a simulated model (step 2), and the ability of the winning model to outperform rival models at predicting 'out-of-sample' data (step 5), can be sufficient to guide model selection. In fact, although these approaches do not allow a specific computational strategy to be linked to a behavioral effect of interest (which remains to be identified), they nevertheless ensure that model comparison and selection are unbiased [9]. To provide a concrete example of this type of studies, we consider the case of researchers interested in determining whether humans preferentially learn from positive compared to negative feedback [25]. This cognitive question may precede the identification of a precise behavioral effect of this learning bias. In this case, the main finding is computational in nature, and may not require a comparison of model simulations to any behavioral effect of interest, once it is established that the model and its parameters can be accurately recovered from a simulated dataset. Note that although model falsification (i.e., the comparison between simulated and observed data) may be skipped in this case, model simulations are still crucial to perform model and parameter recovery.

Finally, it is worth noting that not all model-based studies require a model selection procedure. For instance, when a model is used as an analytical tool rather than being investigated as a computational hypothesis about a cognitive process. Similarly, model selection may not be necessary when the aim is to simply fit a previously well-documented model to investigate the variations of parameter values over individuals or conditions. In these cases, studies should explicitly state the epistemological and methodological status of the model in their analysis plan (Box 1).

Concluding Remarks

In empirical sciences, model-free approaches directly investigate the natural phenomenon of interest, whereas model-based approaches investigate abstract (mathematical) representations of the natural system that are responsible for the empirical phenomenon of interest [26]. The ability to reproduce the empirical phenomenon is therefore crucial to accepting a model as an accurate description of the underlying natural system. From this perspective, we argued that relative model comparison results alone should not be viewed as definitive evidence in favor of a given model explaining any specific phenomenon. In cognitive sciences, in particular, computational models help us to understand the mechanisms that govern thoughts and behaviors through the generation of plausible algorithms that reproduce a behavioral effect of interest [6,7]. In these efforts, computational models are intended not only to predict the observed behaviors but also to describe the cognitive processes that generate the behavioral effect of

Outstanding Questions

How can neural data be used to guide model selection (e.g., testing model predictions at the neurobiological level)?

How can model complexity be assessed beyond the number of parameters (e.g., the functional or architectural complexity of a model)?

How can an adequate model space be defined?

interest. Careful consideration of the recent literature problematically reveals that the 'generative performance' of proposed models has been largely overlooked in favor of their 'predictive performance'. The theoretical argumentation and simulations presented here show that current practice may have led to invalid conclusions (Box 2). The methodological guidelines proposed here should help to amend the current 'state of the art' and increase the explanatory power of computational modeling in cognitive sciences.

Acknowledgments

S.P. is supported by an ATIP-Avenir starting grant (R16069JS) and by a Collaborative Research in Computational Neuroscience ANR-NSF grant (ANR-16-NEUC-0004). E.K. was supported by an advanced research grant from the European Research Council (ERC-2009-AdG-250106). V.W. is supported by a junior researcher grant from the French National Research Agency (ANR-14-CE13-0028). The Institut d'Étude de la Cognition is supported by the LabEx IEC (ANR-10-LABX-0087 IEC) and the IDEX Paris Sciences et Lettres (ANR-10-IDEX-0001-02 PSL*). We thank Charles Findling, and Mehdi Khamassi and Joaquin Navajas for useful comments on an earlier version of the manuscript.

References

1. Montague, P.R. *et al.* (2012) Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80
2. O'Doherty, J.P. *et al.* (2007) Model-based fMRI and its application to reward learning and decision making. *Ann. N. Y. Acad. Sci.* 1104, 35–53
3. Lee, D. (2013) Decision making: from neuroscience to psychiatry. *Neuron* 78, 233–248
4. Forstmann, B.U. *et al.* (2011) Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends Cogn. Sci.* 15, 272–279
5. Maia, T.V. and Frank, M.J. (2011) From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* 14, 154–162
6. Marr, D. (1982) *Vision: A Computational Investigation of Visual Representation in Man,* Henry Holt and Co.
7. O'Reilly, R.C. and Munakata, Y. (2000) Computational explorations in cognitive neuroscience. *J. Math. Psychol.* 46, 504
8. Pitt, M.A. and Myung, I.J. (2002) When a good fit can be bad. *Trends Cogn. Sci.* 6, 421–425
9. Corrado, G.S. *et al.* (2009) The trouble with choice: studying decision variables in the brain. In *Neuroeconomics: Decision Making and the Brain* (Glimcher, P.W. *et al.*, eds), pp. 463–480, Academic Press
10. Daunizeau, J. *et al.* (2014) VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput. Biol.* 10, e1003441
11. Busemeyer, J.R. and Diederich, A. (2014) Estimation and testing of computational psychological models. In *Neuroeconomics: Decision Making and the Brain* (2nd edn) (Glimcher, P. and Fehr, E., eds), pp. 49–61, Academic Press
12. Dienes, Z. (2008) *Understanding Psychology As a Science: An Introduction to Scientific and Statistical Inference,* Palgrave Macmillan
13. Popper, K.R. (1959) *The Logic of Scientific Discovery,* Hutchinson & Co.
14. Platt, J.R. (1964) Strong inference. *Science* 146, 347–353
15. Steingroever, H. *et al.* (2014) Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision* 1, 161–183
16. Drugowitsch, J. *et al.* (2016) Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* 92, 1398–1411
17. Palminteri, S. *et al.* (2015) Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* 6, 8096
18. Ahn, W.-K. *et al.* (2008) Comparison of decision learning models using the generalization criterion method. *Cogn. Sci.* 32, 1376–1402
19. Palminteri, S. *et al.* (2016) The computational development of reinforcement learning during adolescence. *PLoS Comput. Biol.* 12, e1004953
20. Collins, A. and Koechlin, E. (2012) Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol.* 10, e1001293
21. Pitt, M.A. *et al.* (2006) Global model analysis by parameter space partitioning. *Psychol. Rev.* 113, 57–83
22. Viejo, G. *et al.* (2015) Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Front. Behav. Neurosci.* 9, 225
23. Daw, N.D.D. *et al.* (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215
24. Donoso, M. *et al.* (2014) Foundations of human reasoning in the prefrontal cortex. *Science* 344, 1481–1486
25. Lefebvre, G. *et al.* (2017) Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* 1, 0067
26. Weisberg, M. (2007) Who is a modeler? *Br. J. Philos. Sci.* 58, 207–233
27. Palminteri, S. *et al.* (2012) Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* 76, 998–1009
28. Hull, C.L. (1943) *Principles of Behavior: An Introduction to Behavior Theory,* Appleton-Century
29. Rescorla, R.A. and Wagner, A.R. (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (Black, A.H. and Prokasy, W.F., eds), Appleton Century Crofts
30. Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138
31. Padoa-Schioppa, C. (2013) Neuronal origins of choice variability in economic decisions. *Neuron* 80, 1322–1336
32. Watkins, C.J.C.H. and Dayan, P. (1992) Q-learning. *Mach. Learn.* 8, 279–292