HHS Public Access

Author manuscript

Curr Opin Behav Sci. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

Curr Opin Behav Sci. 2016 October ; 11: 49–54. doi:10.1016/j.cobeha.2016.04.003.

Taming the beast: extracting generalizable knowledge from computational models of cognition

Matthew R. Nassar and **Michael J. Frank**

Department of Cognitive, Linguistic and Psychological Sciences, Brown Institute for Brain Science, Brown University, Providence RI 02912-1821

Abstract

Generalizing knowledge from experimental data requires constructing theories capable of explaining observations and extending beyond them. Computational modeling offers formal quantitative methods for generating and testing theories of cognition and neural processing. These techniques can be used to extract general principles from specific experimental measurements, but introduce dangers inherent to theory: model-based analyses are conditioned on a set of fixed assumptions that impact the interpretations of experimental data. When these conditions are not met, model-based results can be misleading or biased. Recent work in computational modeling has highlighted the implications of this problem and developed new methods for minimizing its negative impact. Here we discuss the issues that arise when data is interpreted through models and strategies for avoiding misinterpretation of data through model fitting.

Behavioral and physiological data in systems and cognitive neuroscience are generally collected in reduced environments and constrained experimental conditions, often designed to be diagnostic of competing theories. The generalization of knowledge from simple experiments is crucial to advance our broader understanding of brain and behavior. However, interpreting data according to existing theories causes our knowledge to be conditioned on the quality of said theories and the assumptions thereof. In many cases these conditions are met and theory can drive scientific progress by reducing a dazzling array of neuroscientific data into simpler terms, yielding falsifiable predictions. In other cases a general overarching theory can lead to wasted resources and, at worst, can even impede scientific progress.

Both the advantages and potential dangers of theory are amplified for computational theories, which provide extremely explicit predictions under a specific set of assumptions. Such theories offer an advantage over more abstract ones in that they make predictions about behavior or neurophysiology that are testable, falsifiable, and comparable across models. They do so by formalizing the fundamental definition of the model and linking it to experimental data through a set of assumptions (e.g., the particular form of the behavioral or neural likelihood distribution, conditional independencies in choice behavior, parameter

M.J.F. is a consultant for Hoffman La Roche pharmaceuticals using computational psychiatry methods.

stationarity, etc.). These assumptions can affect how we interpret evidence for or against a model, how we explain differences in the behavior of individuals or how we ascribe functional roles to biological systems based on physiological measurements. Here we examine the theoretical and practical consequences of these assumptions, paying special attention to the factors that determine whether a given assumption will give rise to spurious interpretations of experimental data. In particular, we highlight and evaluate methods used to minimize the impact of modeling assumptions on interpretations of experimental results.

What can go wrong?

While we appreciate that assumptions implicit in experimental designs are critical for data interpretation, and that inappropriate model selection or fitting techniques can produce misleading results, here we focus specifically on issues that can arise when computational models are appropriately and quantitatively linked to meaningful empirical data [1–4]. Under such conditions, quantitative model fitting offers at least three key advantages: 1) competing models can be compared and ranked according to their abilities to fit empirical data, 2) differences across individuals or task conditions can be assessed according to model parameter estimates, providing potential mechanisms giving rise to those differences, and 3) neural computation can be evaluated by comparing physiological measurements to latent model variables. In this section we discuss recent work that highlights how each of these potential advantages can be negated by unmet assumptions.

The impact of modeling assumptions on the arbitration of competing models was recently highlighted in a factorial comparison of visual working memory models [5]. The study decomposed working memory into three distinct processes (item limitation, precision attribution, and feature binding) to construct a large model-set through exhaustive combination of possible computational instantiations of each process. The model set contained archetypical models such as competing “slots” and “resource” models of working memory capacity limitations, but also contained novel combinations that had never been previously tested. This combinatorial model set was fit to data from ten previously published studies that had come to different conclusions regarding the nature of capacity limitations (roughly half had concluded in favor of discrete slots [6,7], whereas the other half had concluded in favor of continuous resource limitation[8,9]). Despite the fact that the original studies had arrived at contradictory conclusions, rank ordered fits from the combinatorial model set were remarkably consistent across datasets. The contradictory conclusions of the original studies were possible because each study compared only a small subset of models by making fixed assumptions (that differed across study) regarding untested model processes, allowing essentially the same data to be used to support competing and mutually exclusive psychological theories.

Similar issues can arise when estimating parameters within a single model. Computational modeling provides a powerful framework for inferring individual differences in latent parameters governing behavior, such as the learning rate used to incorporate new information in supervised or reinforcement learning tasks [10–13]. However, fairly subtle aspects of model specification can have major effects on this estimation process. One recent computational study showed that the common and seemingly innocuous assumption that

learning rate is fixed over time can have drastic consequences on interpretations of behavioral differences: failure to model adjustments in learning rate led to the spurious conclusion that faster learners were best fit by lower learning rates [14,15]. This is to say that, in at least in some extreme cases, naïve reliance on parameter fits can give rise to an interpretation that is exactly opposite to the truth. A similar phenomenon has been noted in computational studies of reinforcement learning and working memory contributions to behavioral learning: failure to appropriately model the working memory process can lead to spurious identification of deficits in the reinforcement learning system [16,17].

The impact that inappropriate assumptions can have on model selection and parameter estimation in turn corrupts the latent variables that are used to test theories of neural computation. Without valid computational targets, analysis of physiological measurements such as fMRI BOLD or EEG are more likely to yield null results, or worse, provide misleading information regarding the computational origins of behavior [18,19]. Or, more simply put, if we do not have a clear understanding of the computations that govern behavior, what hope do we have of discovering the neural instantiations of those computations?

Avoiding the pitfalls

So how can computational models be used to generalize knowledge across cognitive tasks, contexts and species without falling prey to the risks described above? A prominent notion in statistics is that robust inference can be achieved through iteration of steps comprised of model estimation and model criticism [20]. In formal terms, the estimation step involves estimating parameters that allow for the best description of behavior given a candidate model, and the criticism step involves estimating the probability of all possible datasets that could have been observed given that model parameterization [20,21]. This type of criticism is referred to as predictive checking and allows the modeler to ask whether the empirical data that were observed were “typical” for a given model. If the observed data are atypical for the fit model, then model revisions are necessary.

In practice, researchers are generally focused on particular meaningful features of the data motivated by the experimental design, and the typicality of data is often assessed through analyses designed to probe these key features. Specifically, parameters are estimated through model fitting and then parameterized models are used to simulate data that are subjected to the same descriptive analyses originally applied to the empirical data (e.g., a learning study might be concerned with learning curves and/or asymptotic accuracy in one condition compared to another). This approach depends critically on the precision with which behavior can be characterized through these descriptive analyses: the more precisely a behavior of interest can be quantified through a descriptive analysis, the more diagnostic it will be of model sufficiency [15]. In some cases, ability of simulated data to reproduce basic properties of the original dataset (e.g., conditional accuracy and reaction time) can provide rich information regarding why a given model fits the data better than other candidates [16,22]. In other cases, the failure to adequately describe these key features, or some aspect of them, can reveal an inappropriate assumption or a missing component of the empirical data (see figure 1). For example, distributional analyses of response times can reveal when a model

captures empirical data and where it may miss (e.g., the tail or leading edge of the distribution and how they do or don't differ between task conditions); [23,24], patterns that can be revealed via posterior predictive checks [25,26]. In reward learning tasks, sequential choice behavior can be described as a linear function of outcome and reward history, which allows validation of (or reveals deviations from) the specific patterns of data expected by basic reinforcement learning models [27,28]. This type of analysis was recently extended to directly test a model of how rewards can be misattributed in lesioned monkeys, in a manner that was relatively insensitive to modeling assumptions [29].

A related strategy for evaluating model sufficiency is through diagnostic techniques based on predicted likelihood functions. There is a rich statistical literature on the problems that can arise when residuals are not distributed according to the expected likelihood function [30]. Cognitive computational models can fall prey to similar issues; non-uniformity or heteroscedasticity of residuals can inflate some measures of goodness-of-fit and give undue influence to particular data points leading to increased variance or even bias in parameter estimates [31,32]. In some cases, differences in the model-predicted and actual likelihood functions can be observed directly by computing residuals and systematic mismatches between the two can be corrected through changes to the existing model [33]. The appropriate likelihood function is particularly important if one considers the possibility that some small proportion of trials are generated by alternative processes other than the model in question. For example, in value-based decision making, the commonly used *softmax* choice function assumes that choices are increasingly noisy when the differences between alternative choice values are small, and more deterministic when these value differences grow. There is ample evidence for this type of choice function [34], as opposed to alternative *epsilon-greedy* choice functions in which the level of choice stochasticity is uniform across value differences. But several studies have shown that choice likelihood can deviate substantially from either of these functions (for example as a result of just a few choices driven by a different process altogether, such as attentional lapses), and the failure to take into account this irreducible noise can over- or under-emphasize particular data points and potentially bias results (see figure 2) [17,35–37].

So if a model can simulate the key descriptive findings and links computational variables to experimental data through a suitable likelihood function can concerns regarding untested assumptions be set aside? Not necessarily. While these are important criteria for evaluating inferences drawn from a computational theory, they are not typically exclusive to a single model. Thus the question always remains: could the data be better explained through a different set of mechanisms under a different set of assumptions? While this problem is endemic to all of science and not just computational models, one approach to answering this question is to explicitly validate the robustness of model-based findings across a broad range of assumptions. Typically, such validations are conducted using parameters that were originally fixed for simplicity without strong *a priori* rationale [31,38]. A related strategy for assessing the impact of faulty assumptions is to simulate data from models that make a range of assumptions and attempt to recover information from these models with a separate set of models containing mismatched assumptions. In some cases this strategy has revealed problems, such as with the interpretability of softmax temperature in learning models [15], but in other cases it has highlighted the robustness of specific model-based strategies, such

as in the estimation of fMRI prediction error signals using reinforcement learning models under certain conditions [39].

An alternative to explicitly testing the assumptions through which data are linked to a model is to derive and test axioms that should hold irrespective of assumptions [40]. Just as Euclidean geometry postulates that all triangles should have interior angles that sum to 180 degrees, the equations defining a computational model can often be rearranged to identify sets of equalities or inequalities that the entire class of models must obey. One notable example of this strategy comes from economics, where choice consistency, which was established as a fundamental prediction of utility maximization theory, was pivotal for both the falsification of the theory and the subsequent development of better behavioral models [41–43]. Recently, the same approach has been used to test the suitability of reward prediction error models for description of fMRI, electrophysiology, and voltammetry signals [44–47]. While axioms are not mathematically tractable for all models, the basic approach can be followed by identifying testable predictions or boundary conditions through model simulation [48,49]. On its face, the axiomatic approach seems to differ from those discussed above in philosophy: quantitative model fitting promotes inductive reasoning based on reverse inference across a fixed model space, whereas the axiomatic approach lends itself to rejection and revision of failed explanations.

However, this characterization of standard modeling approaches in cognitive science is missing the concept of criticism. Fitting assumptive computational models allows us to induce knowledge regarding the general structure of our world based on specific examples, but rigorous criticism ensures that the knowledge we gain in this way will generalize outside of our model set and experimental space [20,21,50]. Falsification of specific model predictions guides model revisions that make theories more robust and reduces the likelihood of misleading interpretations of experimental results. Through this lens axiomatic methods can be considered as a specific form of criticism tailored to the core assertions of a computational model. It has been noted that the process of careful model criticism can be thought of as obeying an epistemological theory of hypothetico-deductivism, whereby information is gained through rejecting unlikely models, as opposed to through producing support for more likely ones [51].

While we essentially agree with this perspective, we believe that well specified computational theories and the models that instantiate them provide the best of both worlds in terms of philosophy of science: 1) the ability to induce general knowledge from specific data points within a constrained modeling space [52] and 2) the ability to test, reject, and improve upon existing models through a deductive hypothesis testing approach [53]. A balance of these two approaches should allow steady scientific progress through inductive reasoning kept in check by a commitment to falsification of invalid or inappropriate assumptions.

BIBLIOGRAPHY

1. Daw ND. Trial-by-trial data analysis using computational models. Decision making. 2011 This instructive chapter describes some challenges in fitting computational models to experimental data and practical solutions to these problems. A must read for computational psychology or

neuroscience trainees who are hoping to extract information from complex datasets using computational models.

2. Heathcote A, Brown SD, Wagenmakers EJ. An introduction to good practices in cognitive modeling. An introduction to model-based cognitive neuroscience. 2015
3. Pitt MA, Myung IJ. When a good fit can be bad. Trends in Cognitive Sciences. 2002; 6:421–425. [PubMed: 12413575]
4. de Hollander G, Forstmann BU, Brown SD. Different Ways of Linking Behavioral and Neural Data via Computational Cognitive Models. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2016; 1:101–109.
5. van den Berg R, Awh E, Ma WJ. Factorial comparison of working memory models. Psychological Review. 2014; 121:124–149. This paper develops a component-based comparison of visual working memory models by combining computational elements from different models. They fit data from several previous studies with the factorial model set to show, amongst other things, that the rank ordering of model fits is remarkably consistent and that inconsistencies in interpretations of these datasets were due largely to testing a limited subset of models. [PubMed: 24490791]
6. Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. Nature. 2008; 453:233–235. [PubMed: 18385672]
7. Luck SJ, Vogel EK. Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in Cognitive Sciences. 2013; 17:391–400. [PubMed: 23850263]
8. Bays PM, Husain M. Dynamic shifts of limited working memory resources in human vision. Science. 2008; 321:851–854. [PubMed: 18687968]
9. Ma WJ, Husain M, Bays PM. Changing concepts of working memory. Nature Neuroscience. 2014; 17:347–356. [PubMed: 24569831]
10. Wiecki TV, Poland J, Frank MJ. Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry: Clustering and Classification. Clinical Psychological Science. 2015; 3:378–399.
11. Mars RB, Shea NJ, Kolling N, Rushworth MFS. Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. The Quarterly Journal of Experimental Psychology. 2012; 65:252–267. [PubMed: 20437297]
12. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nature Neuroscience. 2007; 10:1214–1221. [PubMed: 17676057]
13. Cohen MX. Individual differences and the neural representations of reward expectation and reward prediction error. Social Cognitive and Affective Neuroscience. 2006; 2:20–30. [PubMed: 17710118]
14. Nassar MR, Wilson RC, Heasly B, Gold JI. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. Journal of Neuroscience. 2010; 30:12366–12378. [PubMed: 20844132]
15. Nassar MR, Gold JI. A healthy fear of the unknown: perspectives on the interpretation of parameter fits from computational models in neuroscience. PLoS Comput Biol. 2013; 9:e1003015. This paper highlights the dependency of parameter fits on the model in which they are embedded by simulating data from a wide range of learning models and fitting it with simpler “out-of-the-box” learning algorithms. [PubMed: 23592963]
16. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012; 35:1024–1035. [PubMed: 22487033]
17. Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ. Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. Journal of Neuroscience. 2014; 34:13747–13756. [PubMed: 25297101]
18. Ioannidis JPA. Why Most Published Research Findings Are False. Plos Med. 2005; 2:e124. [PubMed: 16060722]
19. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science. 2014; 9:641–651. [PubMed: 26186114]

20. Box G. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society Series A (General). 1980; 143:383–430.
21. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica. 1996 [no volume].
22. Ding L, Gold JI. Separate, Causal Roles of the Caudate in Saccadic Choice and Executionin a Perceptual Decision Task. Neuron. 2012; 75:865–874. [PubMed: 22958826]
23. Ratcliff R, Frank MJ. Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. Neural Comput. 2012; 24:1186–1229. [PubMed: 22295983]
24. Ratcliff R, Smith PL. Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. Journal of Experimental Psychology: General. 2010; 139:70–94. [PubMed: 20121313]
25. Cavanagh JF, Wiecki TV, Kochar A, Frank MJ. Eye Tracking and Pupillometry Are Indicators of Dissociable Latent Decision Processes. Journal of Experimental Psychology: General. 2014; doi: 10.1037/a0035813
26. Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D. fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. Journal of Neuroscience. 2015; 35:485–494. [PubMed: 25589744]
27. Lau B, Glimcher PW. Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys. J Exp Anal Behav. 2005; 84:555–579. This paper examines time-series dependencies in choice behavior by developing regression-based analyses of choice-history. The analysis techniques developed have been crucial for identifying correspondences and differences between human and animal learning behavior and that of computational algorithms designed to simulate it. [PubMed: 16596980]
28. Seo H, Cai X, Donahue CH, Lee D. Neural correlates of strategic reasoning during competitive games. Science. 2014; 346:340–343. This paper identifies differences in the behavior of monkeys and reinforcement learning models using choice history analysis. The authors show that differences in behavior were attributable to monkeys adopting strategic choice behavior to exploit the algorithm employed by a computerized opponent. [PubMed: 25236468]
29. Walton ME, Behrens TEJ, Buckley MJ, Rudebeck PH, Rushworth MFS. Separable Learning Systems in the Macaque Brain and the Role of OrbitofrontalCortex in Contingent Learning. Neuron. 2010; 65:927–939. This paper explores the effects of orbitofrontal lesions on learning and choice behavior using a combination of model-fitting and descriptive statistics. A key methodological innovation was the development of an informative model-free analysis to test the key predictions made by a specific computational deficiency (namely in precisely attributing rewards to relevant actions). [PubMed: 20346766]
30. Belsley, DA.; Kuh, E.; Welsch, RE. Regression Diagnostics. John Wiley & Sons; 2005.
31. Nassar MR, Bruckner R, Gold JI, Li S-C, Heekeren HR, Eppinger B. Knowing what you don't know: age-related deficits in representing uncertainty. Nat Commun.
32. Wiecki TV, Sofer I, Frank MJ. HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Front Neuroinform. 2013; 7:14. [PubMed: 23935581]
33. van den Berg R, Shin H, Chou W-C, George R, Ma WJ. Variability in encoding precision accounts for visual short-term memory limitations. Proceedings of the National Academy of Sciences. 2012; 109:8780–8785.
34. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans [Internet]. Nature. 2006; 441:876–879. [PubMed: 16778890]
35. Collins A, Koechlin E. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. PLoS Biol. 2012; 10:e1001293. [PubMed: 22479152]
36. Frydman C, Camerer C, Bossaerts P, Rangel A. MAOA-L carriers are better at making optimal financial decisions under risk. Proceedings of the Royal Society B: Biological Sciences. 2011; 278:2053–2059. [PubMed: 21147794]
37. Doya K. Modulators of decision making. Nature Neuroscience. 2008; 11:410–416. [PubMed: 18368048]

38. Badre D, Doll BB, Long NM, Frank MJ. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron. 2012; 73:595–607. [PubMed: 22325209]
39. Wilson RC, Niv Y. Is Model Fitting Necessary for Model-Based fMRI? PLoS Comput Biol. 2015; 11:e1004237. This paper tests whether identification of an fMRI prediction error signal depends on accurate parameter estimation. The authors use a closed-form expression for the relationship between simulated and recovered fMRI signals over a wide range of parameterizations to show that fMRI analyses are relatively robust to exact parameterization. [PubMed: 26086934]
40. Caplin A, Dean M. Axiomatic methods, dopamine and reward prediction error. Current Opinion in Neurobiology. 2008; 18:197–202. [PubMed: 18678251]
41. Samuelson PA. A note on the pure theory of consumer's behaviour. Economica. 1938
42. Savage, LJ. The Foundations of Statistics. Wiley Publications in Statistics; 1954.
43. Ellsberg D. Risk, ambiguity, and the Savage axioms. The Quarterly Journal of Economics. 1961 [no volume].
44. Caplin A, Dean M. Dopamine, reward prediction error, and economics. The Quarterly Journal of Economics. 2008 [no volume].
45. Rutledge RB, Dean M, Caplin A, Glimcher PW. Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. Journal of Neuroscience. 2010; 30:13525–13536. [PubMed: 20926678]
46. Hart AS, Rutledge RB, Glimcher PW, Phillips PEM. Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term. Journal of Neuroscience. 2014; 34:698–704. This paper shows that dopamine levels in the nucleus accumbens core satisfy a set of axioms defining the minimum criteria for encoding a reward prediction error (RPE) signal. [PubMed: 24431428]
47. Bayer HM, Glimcher PW. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron. 2005; 47:129–141. [PubMed: 15996553]
48. Cockburn J, Collins AGE, Frank MJ. A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice. Neuron. 2014; 83:551–557. [PubMed: 25066083]
49. Busemeyer JR, Townsend JT. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review. 1993; 100:432–459. [PubMed: 8356185]
50. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics. 1984 [no volume].
51. Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol. 2012; 66:8–38. [PubMed: 22364575]
52. Savage LJ. The foundations of statistics reconsidered. 1961
53. Popper K. The Logic of Scientific Discovery. 2005 [no volume].
54. McGuire JT, Nassar MR, Gold JI, Kable JW. Functionally dissociable influences on learning rate in a dynamic environment. Neuron. 2014; 84:870–881. [PubMed: 25459409]

HIGHLIGHTS

- Computational models extract general principles from specific data.
- However, such principles are conditional on modeling assumptions.
- Faulty assumptions can bias findings and mislead data interpretation.
- The modeler’s toolkit includes several techniques to avoid this pitfall.
- These techniques are critical to broadly interpret computational findings.

Box 1

Computational model

A mathematical description of the processes giving rise to behavioral or physiological phenomena that in many cases reflects a tractable, concrete instantiation of a cognitive theory.

Model assumptions

The complete set of transformations through which a computational model generates behavioral or physiological data. This includes both fundamental assumptions, which are critical predictions of the overarching cognitive theory, as well as structural assumptions such as about the distributional form, conditional dependency, or stationarity of variables. While structural assumptions are not generally central to the cognitive theory, they nevertheless have important implications for the testing of cognitive theories using computational models.

Parameter

A quantity of the computational model that is allowed to take a range of values to achieve a range of different behaviors and account for individual differences without requiring an entirely different model. For example, one could think about the volume knob on a radio as controlling a parameter that defines the amplitude of sound waves.

Model fitting

The process of adjusting model parameters to values that maximize some utility function, often log likelihood of the observed data.

Model selection

The process of determining which model within a set provides the best description of the data.

Likelihood function

The function used to directly map internal representations (such as the reward value of possible actions in decision making) onto a probability distribution over observed data (the chosen actions).

Latent variable

An element of a computational model that is internally regulated (ie. does not require external selection or adjustment) to allow for different modes of model behavior. For example, action values are adjusted over time in reinforcement learning models to allow for adaptation to changes in action outcome contingency.

Predictive checking

The practice of simulating data from computational models using parameter values fit to empirical data in order to determine which aspects of empirical data the model can or cannot account for.

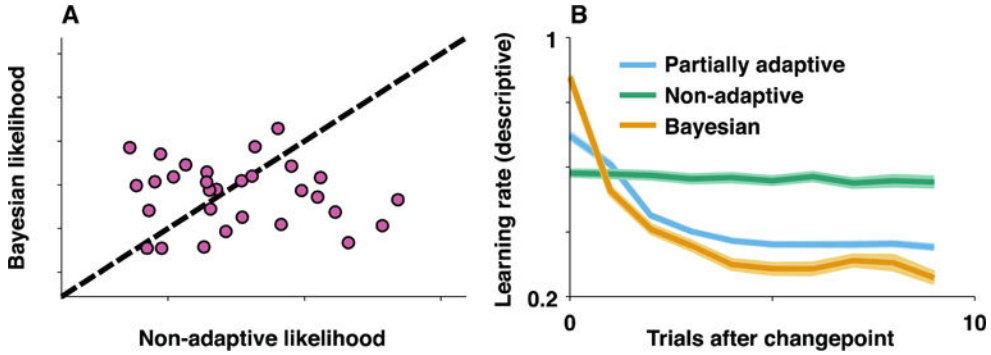

Figure 1. Predictive checking can reveal model insufficiency and guide model improvement

Behavioral data were simulated for 30 subjects in dynamic inference task using an error-driven learning rule that adapts learning rates in accordance with subject behavior (eg. learning rates were adapted in accordance with principles of normative learning, but less so than would be prescribed by an optimal inference algorithm) [14]. Simulated data was fit using two separate models: 1) a Bayesian model assuming rational inference with 2 free parameters (hazard rate, update variability) and 2) a fixed learning rate error-driven learning model with 2 free parameters (learning rate, update variability). **A)** Both models provide a reasonable fit to the data in terms of likelihood and neither model is systematically preferred across simulated subjects. **B)** Simulated data from partially adaptive model shows a characteristic decay in learning rate after task change-points (dark/light blue indicate mean/SEM across simulated subjects). To facilitate predictive checking, data was simulated from Bayesian and fixed learning rate models using maximum likelihood parameter values for each simulated subject. Descriptive analyses were used to separately estimate learning rates for subsets of trials according to how recently the outcome contingencies of the simulated task underwent a fundamental change [54]. Simulated data from Bayesian (yellow) and fixed learning rate (green) models over- and under-estimate learning rate dynamics respectively and thus both fail the predictive checking procedure. In this case the failed predictive check not only reveals model insufficiency, but also sheds light on how the models could be improved (eg. through parametric modulation of learning rate dynamics).

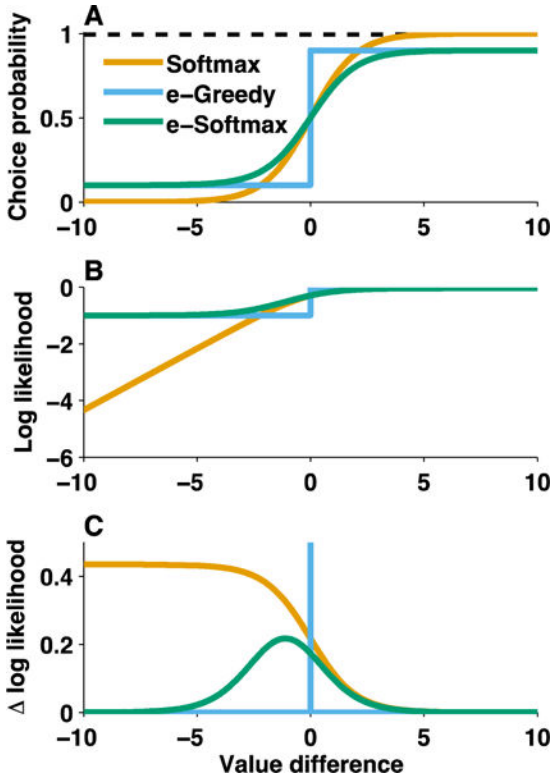

Figure 2. Likelihood function determines which data points are most influential for model fitting. A)

Probability of a rightward choice (ordinate) was computed analytically for an agent performing a two alternative choice task across a range of value differences (abscissa) using three different likelihood functions. While each likelihood function dictates a higher probability of choosing the higher value option, the exact shapes of the functions differ. Softmax (β=1) action selection dictates a sigmoid choice probability function with asymptotes of 0 and 1 for negative and positive value differences respectively (this function assumes increasingly deterministic choice as value differences are large enough). In contrast, epsilon-greedy (e=0.2) action selection dictates a step function with asymptotes at 0.1 and 0.9 (this function assumes exploitation of higher value options with some noise but with no difference in the degree of exploitation as a function of value differences). E-softmax (β=1, e=0.2) is a mixture of these two functions, assuming softmax value-based action selection but with some irreducible noise; it has a sigmoid shape and asymptotes at 0.1 and 0.9. **B)** Model fitting typically involves maximizing log-likelihood of the data for a given model. Considering the log likelihood of a rightward choice for different levels of value difference provides an indicator of the trials that matter most for this process. For the softmax function, unpredicted rightward responses (large negative value difference) are severely penalized in terms of log likelihood, and hence likelihood maximization will adjust model parameters to increase value differences for rightward choices. In contrast, differences in the log likelihood of expected responses (e.g. positive value differences) are deemphasized such that it is difficult to see the difference between asymptotic log likelihoods of softmax [log(1)] and e-greedy [log(1-e)] functions. **C)** Depending on the assumed likelihood function, some rightward shifts will be more influential. The impact of a rightward shift on log likelihood,

or the derivative of the log likelihood function, is plotted at each value difference for each likelihood function. The likelihood functions differ dramatically in sensitivity to different value differences: 1) softmax gives most credit to improvements in value difference of the most unlikely outcomes, 2) e-greedy only cares about the ones near zero (indifference) and 3) e-softmax cares most about pushing slightly negative value-differences toward (or above) zero. A key takeaway is that choosing a likelihood function has fairly strong statistical implications on the types of errors that will have the greatest effects on model selection and parameter estimation.