Model virtues in computational cognitive neuroscience

Heijnen, S.; Sleutels, J.J.M.; Kleijn, R.E. de

Model Virtues in Computational Cognitive Neuroscience

Saskia Heijnen, Jan Sleutels, and Roy de Kleijn

Abstract

■ There is an abundance of computational models in cognitive neuroscience. A framework for what is desirable in a model, what justifies the introduction of a new one, or what makes one better than another is lacking, however. In this article, we examine key qualities ("virtues") that are desirable in computational models, and how these are interrelated. To keep the scope of the article manageable, we focus on the field of cognitive control, where we identified six "model virtues": empirical accuracy, empirical scope, functional analysis, causal detail, biological plausibility, and psychological plausibility. We first illustrate their use in published work on Stroop modeling and then discuss what expert modelers in the field of cognitive control said about them in a series of qualitative interviews. We found that virtues are interrelated and that their value depends on the modeler's goals, in ways that are not typically acknowledged in the literature. We recommend that researchers make the reasons for their modeling choices more explicit in published work. Our work is meant as a first step. Although our focus here is on cognitive control, we hope that our findings will spark discussion of virtues in other fields as well. ■

INTRODUCTION

The range of computational models used in cognitive neuroscience is very wide. Model families include sequential sampling models, connectionist or neural network models, reinforcement learning models, predictive processing or Bayesian models, and dynamical systems models. Within each there is a great variety of model types implementing the principles of the model family. For instance, the family of sequential sampling models includes linear ballistic accumulator models, drift diffusion models, and lognormal race models. In turn, each of these has different implementations depending on the parameters used and the theoretical load attached to them.

For older models, the plurality can partly be understood in terms of Marr's three levels of analysis ("computational theory," "representation and algorithm," and "hardware implementation"; Marr, 1982). In the first 3 decades after the publication of Marr's landmark book, many models explicitly situated themselves at one of the three levels. They explored styles of modeling that were best suited for one particular level of analysis, without bothering much about modeling requirements at other levels (Peebles & Cooper, 2015; Poggio, 2010). As computational neuroscience matured, references to Marr's levels have become very scarce. Newer models typically do not situate themselves at one of Marr's levels, but rather try to straddle all three of them (Collins & Shenhav, 2022; Niv & Langdon, 2016).

Although modeling choices today are no longer directly motivated by Marr's levels of analysis, the plurality remains (Blohm, Kording, & Schrater, 2020). This can to some extent be understood in terms of the different vested interests of researchers and their teams, based on previous training, acquired expertise, different local research traditions, and so on. Such reasons are only contingent, however, in the sense that they do not apply across different research traditions. Moreover, the choice for specific "modeling flavors" remains mostly implicit in the literature. This creates a situation in which research traditions seem to be locked into their own preference bubbles.

For preferring one type of model over another, one would ideally expect researchers to have reasons that can be meaningfully communicated and systematically discussed across traditions. Such considerations are mostly lacking in the literature, however. When reasons are given at all, these tend to offer ad hoc support for favoring one particular new model over an earlier one, rather than offering a systematic reflection on what computational models are expected to do across the board—which qualities they are expected to possess, which purposes they are expected to serve.

We aim to explicate reasons for choosing or designing models, and explore how these reasons interrelate. We seek to identify and discuss some of the main "modeling virtues," as we will call them: key qualities of computational models that are particularly valued. We seek to specify these in a more or less uniform manner, and to make it possible to systematically reflect on how they relate to each other. We thus hope to promote and facilitate discussion between different modeling traditions. To keep the scope of the article manageable, we chose to focus on the field of cognitive control. Our work is meant as a first step, which we hope will spark discussion of virtues in other fields as well.

Leiden University

Journal of Cognitive Neuroscience 36:8, pp. 1683–1694
https://doi.org/10.1162/jocn_a_02183

In the Model Virtues section, we present six virtues that stand out when reflecting on the literature and on a series of qualitative interviews with expert modelers in the field of cognitive control, who were invited to comment on the role of virtues in their work. An Example: Modeling the Stroop Effect discusses examples of these virtues in empirical literature. We then present the results of the qualitative interviews. Although there is convergence of opinion between some of the respondents, there is also disagreement, which means that there is room for discussion about the role played by modeling virtues once they have been made explicit and systematized.

Table 1. Three Key Aspects of Computational Models in Cognitive Neuroscience, and Six Corresponding Model Virtues

Aspects	*Virtues*
Empirical adequacy	Empirical accuracy
	Empirical scope
Explanatory power	Functional analysis
	Causal detail
Interpretability	Biological plausibility
	Psychological plausibility

MODEL VIRTUES

In philosophy of science, there has been much discussion about theory choice: Why prefer one theory above another? Which qualities of a theory are valued in particular? In the literature, these qualities have come to be referred to as "theoretical virtues" (e.g., Keas, 2018; Schindler, 2018). Analogously, the literature on "explanatory virtues" focuses on what makes a particular explanation preferable over competitors (e.g., Rosales & Morton, 2021; Mackonis, 2013; Ylikoski & Kuorikoski, 2010).

There is no similar discussion with regard to models. This is a pertinent matter, because models arguably have taken the place of theory in cognitive science (Miletić, Boag, & Forstmann, 2020; Sun, 2009; Newell, 1990), and most explanation in this field is model based (Lawler & Sullivan, 2021; Bokulich, 2017). Considering the central role of models in cognitive neuroscience, the lack of discussion regarding model virtues is a regretful lacuna.

Taking our lead from the literature on theoretical and explanatory virtues, we investigated whether similar considerations are at play in computational modeling. In published work, one finds only incidental hints at motivations for working with a particular type of model. In a series of qualitative interviews, we conducted with expert modelers; however, we found that virtues do indeed play an important role behind the scenes, much more than is reflected in published work. The interviews yielded no less than 25 qualities that researchers consider when developing and evaluating computational models. Closer analysis of the responses showed considerable overlap between these qualities, however, with different researchers tending to use different terminology to capture what is essentially the same relevant aspect or virtue of computational models. We identified three relevant aspects and six corresponding virtues, which we used for clustering the results from the interviews (for details, see Appendix B).

The resulting framework is shown in Table 1. We will use this framework to discuss the relations of mutual support and trade-off between the different virtues, and to illustrate the importance of a more systematic discussion of relations between virtues.

Empirical adequacy is the extent to which a model matches the data. A model can "match the data" in at least two different (and mutually independent) ways, which we use here to define the virtues of empirical accuracy and empirical range. *Empirical accuracy* is higher (compared with another model) as the model's predicted value is closer to the observed value, or as the predicted confidence interval centered around the observed value is smaller. The virtue of *empirical scope* is satisfied to a higher degree (compared with another model) as a wider range of data sets or effects can be reproduced by a model (to a certain degree of precision).

Explanatory power is a model's ability to help us understand the target phenomenon. This too can be achieved in at least two different (and arguably independent) ways, which we use here to define the virtues of functional analysis and causal detail. *Functional analysis* is the extent to which a model analyzes the target system or process in terms of its component computational resources, that is, identifies the system's functional architecture at an appropriate level of aggregation (Pylyshyn, 1984). A model can give functional analysis while still containing what critics will call "black boxes." The virtue of *causal detail* targets these black boxes and refers to the degree of detail with which we understand how the components work and interact.

Finally, *interpretability* is a model's susceptibility to meaningful interpretation. This too can be achieved in at least two different (and possibly independent) ways, which we use here to define the virtues of biological plausibility and psychological plausibility. *Biological plausibility* is satisfied to a higher degree as a model's parameters are more readily related to biological components (at a specified level of abstraction) that are taken to be responsible for the target phenomenon. A model with biological plausibility is an abstracted rendition of the target system from a biological point of view. Notice that all models are bound to gloss over numerous biological details, which means that biological plausibility is always keyed to a specific level of detail and brings with it a certain level of abstraction.

Psychological plausibility is satisfied to a higher degree as the parameters of a model are more readily related to functional components (at a specified level of abstraction) that are taken to be responsible for the target phenomenon. Analogous to biological plausibility, a model with

psychological plausibility is an abstracted rendition of the target system from a cognitive-functional point of view.

Notice that correlations between some of these virtues are to be expected. For example, causal detail in cognitive neuroscience is likely to be cashed out in biological terms. Because biological plausibility is keyed to a certain level of abstraction, however, it does not necessarily enhance causal detail. Another example is functional analysis, which usually (but not necessarily) coincides with psychological plausibility. If a computational model lays out the functional components that make up the target phenomenon, it will typically do so by identifying these components in terms of concepts taken from cognitive psychology, thereby enhancing the model's cognitive salience, that is, the model's ability to cut cognitive processes at their joints.

Before turning to the results from the interviews, in the next section, we first illustrate how the key virtues defined here can guide specific modeling choices. As our example, we chose computational modeling of the Stroop effect, a widely studied phenomenon for which many different models have been proposed.

AN EXAMPLE: MODELING THE STROOP EFFECT

In the Stroop task, participants are asked to name the color of a given stimulus, whereas the stimulus itself is a word that names a color (Stroop, 1935). The delay in RT between naming the color of a stimulus that spells a different name (incongruent trials) and naming the color of a stimulus that spells the same name (congruent trials) is called the "Stroop effect." The Stroop task spawned countless studies of this and related effects, as well as explanations in terms of many different types of models (Macleod, 1991), which makes it a great example to study virtues in action. Please refer to Text Box 1 for more information on the Stroop models discussed here.

Box 1: Computational models of the Stroop task.

The *Stroop task* is a well-known psychological test measuring the delay in RT between congruent and incongruent color words (the *Stroop effect*). Stroop's (1935) original article is one of the most cited articles in experimental psychology, and there has been a wide range of models trying to explain its mechanism. This box briefly explains the main classes of models.

Connectionist models model the mechanism as a network, with the activity flowing between nodes representing neural or cognitive content. An early example is Cohen, Dunbar, and McClelland (1990), who developed a connectionist model in which an attentional mechanism (not specified) acts to modify the strengths of the pathways in the model. There are three groups of input units, representing ink color, word identity, and task demand (color naming or word reading). In the next layer, there are nodes in which task demand modulates the activity fed forward from the input layer, such that the pathway with the intended task is more easily activated. The result is fed to the output layer, which has "red" and "green" as responses. Botvinick, Braver, Barch, Carter, and Cohen (2001) add a single unit to a successor of this model: a conflict-monitoring unit to detect when processes are in conflict with each other (word reading vs. color naming) and then allocates attention to prior goals (task demand). Conflict occurs when two units that inhibit each other's activity are simultaneously activated.

Reinforcement learning approaches model the mechanism with online learning resulting from the participant's actions. Shenhav, Botvinick, and Cohen (2013) add a unit that implements an optimization strategy—the expected value of control. This is a function of (1) the payoff expected from a controlled process, (2) the amount of control that must be invested to achieve that payoff, and (3) the cost in terms of cognitive effort. Lieder, Shenhav, Musslick, and Griffiths (2018) expand on the previous models by positing a reinforcement learning algorithm that learns the value of control, and adding Bayesian rules for determining the optimal weights in the model.

Sequential sampling models explain the effect through one or more accumulators moving toward a decision threshold. Fennell and Ratcliff (2019) present a model in which evidence for one option (e.g., red) is evidence against the others (e.g., blue, green). The decision-making process, they argue, is prominently determined by drift rate (a measure of the quality of evidence of the stimulus) and by the decision boundary (reflecting differences in response styles that vary across individuals), which are both cognitively interpretable parameters in the model they propose. Three further parameters account for the assumption that the cognitive system processes things slightly differently every time: across-trial variability in drift rate, in decision boundaries, and in nondecision time. Finally, the model has a parameter for fluctuations in evidence accumulation within each trial.

Mathematical models do not attempt to explain the mechanism itself but describe the resulting RT distribution in terms of different parameterizations. Heathcote, Popiel, and Mewhort (1991) critique the use of mean RT in analyses, because it causes a loss of information about performance. Instead, the shape of RT distributions should be taken into account. Hence, they propose an ex-Gaussian model that fits to the shape of the RT distribution using

parameters for its mean, variance, and the parameter of the exponential component. As it is a mathematical model, the authors do not associate the elements of the model to cognitive or neural processes.

Bayesian models of the Stroop effect incorporate measures of uncertainty and focus on cognitive control as statistical inference. The Bayesian algorithms used behave similarly to reinforcement learning algorithms with varying learning rates. Although not attempting to explain the Stroop effect per se, instead of focusing on cognitive control, Jiang, Heller, and Egner (2014) propose such a hierarchical Bayesian model that generates estimates of volatility and conflict, and then secondly updates these estimates given what was observed. The model then gives a probability distribution over the predicted conflict level variable. With this model, they want to account for the flexibility of cognitive control: The belief about the volatility of the environment determines the effect that longer-term and short-term events have on future predictions, with a more stable environment allowing longer-term events to have a stronger effect.

Empirical Accuracy in Stroop Models

Empirical accuracy was defined in terms of a model's fit to a particular data set: Empirical accuracy increases as the model predicts a value that is closer to the observed value, or a smaller confidence interval centered around the observed value. Lieder and colleagues (2018) exemplify this when pitting their learned value of control model against two alternatives, the former winning in terms of the Bayesian Information Criterion (p. 16).

An appreciation of "mere" empirical accuracy can be found in Heathcote and colleagues (1991). They propose an ex-Gaussian model to address interference and facilitation effects in Stroop RT distributions. They point out that the merit of their model is that it provides a good *description* of the data, although it is not susceptible of a psychological interpretation (pp. 346–347).

Empirical Scope in Stroop Models

The virtue of empirical scope focuses not on a model's fit with one particular data set, but on a model's ability to reproduce multiple data sets and observed effects. This virtue is commonly applied in the literature to support the introduction of a new model. For example, Cohen and colleagues (1990), Jiang and colleagues (2014), and Chuderski and Smolen (2016) all mention a particular effect, or a set of effects, that is not captured by previous models, and propose a model that does reproduce these empirical observations.

Functional Analysis and Psychological Plausibility in Stroop Models

As mentioned earlier, the virtues of functional analysis and psychological plausibility often go together in cognitive modeling, where they appear as cognitive salience: a model's ability to dissect a cognitive process into psychologically interpretable components. A good example of this in Stroop modeling is Fennell and Ratcliff's (2019) sequential sampling model for multichoice decision-making (RTCON2). The decision-making process, they argue, is prominently determined by drift rate (a measure of the quality of evidence of the stimulus) and by the decision boundary (reflecting differences in response styles that vary across individuals), which are both parameters in the model they propose. Thus, the new model not only accounts for behavioral data (empirical accuracy) but also has psychologically interpretable parameters that reflect the functional components of decision-making processes (p. 2100).

Causal Detail in Stroop Models

Authors frequently criticize models for containing a black or gray box, which they subsequently set out to unpack in a new model. For instance, Botvinick and colleagues (2001) point out that previous models do not address the question of how the cognitive system determines when control is required, and they then provide an account of this (p. 624). Another example is Shenhav and colleagues (2013), who note that when it comes to the functioning of the dorsal anterior cingulate cortex, a wide range of findings is typically reduced to a single basic computation. This hampers an integrated functional understanding, and they provide a more detailed account (p. 217).

Biological Plausibility in Stroop Models

A model is biologically plausible to the extent that the model's parameters are susceptible to interpretation in terms of particular brain structures and their functions. Examples in Stroop modeling are Botvinick and colleagues (2001) and Shenhav and colleagues (2013), who spend a good part of their articles establishing links between their proposed models and literature on brain structure and function. Similarly, Jiang and colleagues (2014) criticize competing models for positing mechanisms that are unlikely to be true of the brain (p. 40).

Reflection: Virtues in Stroop Models

All of the six key virtues can be found in the literature. They are typically mentioned to provide ad hoc support for favoring one particular model over another. There is no systematic reflection on why one particular virtue of a model is taken into consideration rather than another, which virtues computational models should have across the board, or which trade-offs between virtues are to be expected.

One example of a trade-off between virtues is the fact that biological and psychological plausibility, causal detail, and functional analysis all impose restrictions on the computational model used to fit the data. This is largely seen as a benefit of these virtues: They constrain the space of possibilities. This also means, however, that empirical accuracy tends to be sacrificed. Conversely, a statistical model that does not have these restrictions can fit the data to a higher degree of accuracy, but this typically comes at the cost of cognitive salience and biological plausibility. Note that Niv (2021) argues that a clever behavioral paradigm can offer greater constraints on computational models than neural studies.

As another example, expanding empirical scope typically comes at the cost of empirical accuracy (Verguts, 2022, ch. 5). Similarly, expanding empirical scope tends to decrease the model's level of detail, which means that functional analysis and/or causal detail may have to be sacrificed to expand empirical scope (Lovett, 2002).

Conversely, authors focusing on causal detail typically tend to explicitly limit the intended scope of their model. Note that causal detail has a delicate relationship with understanding: We may get better understanding of specific causal factors for a given cognitive phenomenon, but we lose the bigger-picture understanding that more abstracted or idealized models give us (Potochnik, 2015).

It is important to realize that a virtue is not valuable in and of itself; its value depends on the particular research questions that a model is intended to address, as well as on its relations to other virtues. In the next section, we will see that modelers give more thought to the value and interrelations of virtues than is reflected in the literature.

INTERVIEW RESULTS

Semistructured interviews were held with 15 acknowledged experts in computational modeling of cognitive control. Details about the methods of the interview can be found in Appendix A. Here, we discuss those parts of the interviews that focused on the six virtues introduced above. A full overview of virtues that came out of the interviews is presented in Appendix B. It is interesting to note that none of the interviewees linked their modeling choices to any of Marr's levels in particular and that several explicitly said that models should try to straddle all three of Marr's levels (respondents 2, 6, 7, 11, 12, 14).

We will see that researchers are aware that the value of virtues is context dependent and that trade-offs between them must be taken into consideration. We will also find that researchers agree on some considerations while disagreeing about others, which means that there is room for discussion about the role played by modeling virtues once they have been made explicit and systematized.

Empirical Accuracy in the Interviews

Most respondents said that empirical accuracy is important as arbitrator between competing models (1, 2, 3, 9, 13). One respondent explicitly objected to this, however. Considering that all models, by their very nature, are simplifications and abstractions, it makes no sense to pit one against the other: They are both wrong, and each of them is leaving out things that might change the whole story (5).

Many respondents took a qualified stance on empirical accuracy, indicating that accuracy should be balanced with other virtues, and can even be outweighed by them. Some argued that all models are simplifications and hence should not be expected to match all the data; you want them to match theoretically meaningful aspects of the data (4, 5, 7, 12). In a similar vein, some respondents noted that empirical accuracy sometimes can and should be sacrificed for the sake of abstraction, understanding, and making intuitive sense (6, 8, 15).

Two respondents observed that a good fit does not make a good model (7, 12), and two others pointed out that matching the data is not sufficient for explanation, because it does not provide a mechanism (8) or guarantee understanding (11). One respondent said that no model will ever get it entirely right, so other virtues of the model will have to be weighed against what the model cannot explain (14). Another respondent said that we do not want a model that perfectly fits, because this would yield models as big as the system itself, which defies the purpose of modeling (15). The problem of "overfitting" was mentioned by several respondents (2, 3, 7, 13).

Three respondents indicated that in some cases, they were willing to accept models that did not fit any data at all, for example, when the model only points up trends in the data that indicate qualitative differences (8). Similarly, a model can be highly biophysically realistic, to the extent that there are no available data at that level of detail (12), or a model can be a powerful proof-of-principle (14).

It became clear in the interviews that researchers are much more open to considering trade-offs and other systematic relations between different virtues than is reflected in the literature. There appear to be many reasons not to maximize empirical accuracy.

Empirical Scope in the Interviews

Empirical scope as defined here was sometimes referred to as "generalization" by the respondents: Does a model generalize to other data sets, other tasks, or other types of processes? One respondent pointed out that what is important for a model is not just that it fits a given set of

data but also that it can bring one data set into contact with some new set of data or with another model (4). Most respondents valued a model's capacity to address not just one phenomenon or data set, but others as well (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13). This was said to be a core piece of explanation (1) or the essence of science (5). Another respondent observed that some of the biggest breakthroughs come from generalizing from one domain to another (3). At the same time, a number of respondents acknowledged that this kind of "big picture" work is not common; most work is paradigm bound and focused on specific task domains (1, 3, 6, 12).

Summarizing, respondents were almost unanimous in valuing empirical scope, while at the same time acknowledging that its role in practice is very limited. Surprisingly, none of the respondents mentioned the trade-off between empirical scope, functional analysis and causal detail.

Functional Analysis and Psychological Plausibility in the Interviews

The clustered virtues of functional analysis and psychological plausibility were discussed under a variety of names, including "psychological plausibility," "cognitive salience," "functional understanding," and "insight in the cognitive phenomenon."

Some respondents mentioned psychological plausibility as an important virtue (1, 3, 9, 13). Two of these (3 and 13) stated that a cognitive-level story by itself can yield deep understanding, so that there is merit in models that do not address the biological level at all. They hoped that those levels would eventually come together, but addressing biology is not a requirement for every model, certainly not in the early stages of its development.

Cognitive salience was frequently mentioned in respondents' discussions of the interpretability of a model's parameters. Five respondents said they want to be able to interpret the model's parameters (3, 6, 7, 9, and 14). Two of those said they want to be able to map the features of the model to the phenomenon (6 and 7), whereas another said they value it when parameters have "intrinsic meaning," that is, when they relate to a cognitive process (9). Two respondents said they want to understand "what the model is doing," that is, why is it giving a particular result (3 and 6). Two others did not require this per se, but still wanted the parameters to be psychologically interpretable (7 and 9). The former two, then, would accept a narrower range of models: Only those that have psychologically interpretable parameters and allow one to understand how the model works. They argued that if you are unable to explain why the model can account for the data, the model does not enhance your understanding of the process, and therefore lacks value.

In a related context, many of the respondents observed that a great strength of computational models (as compared with verbal-conceptual or statistical models) is that you can "play" with them: By studying how different parameter settings affect the model's behavior, you enhance your insight in the process, that is, you get deeper understanding of the cognitive phenomenon (2, 3, 4, 5, 6, 11, 12, 13, 14).

Summarizing, respondents generally found it important that a model can be interpreted in a meaningful way, but only some explicitly mentioned psychological interpretability here. Many respondents valued the insight generated through playing with a model, which is closely related to functional analysis.

Causal Detail in the Interviews

The virtue of causal detail (negatively put, a model's lack of black boxes) was discussed under different names in the interviews, most notably including "explanatory completeness," "complexity," and "mechanistic explanation." As mentioned earlier, causal detail in cognitive neuroscience is likely to be realized in abstracted biological terms.

In the context of causal detail, one respondent saw coming up with more precise and more powerful descriptions of brain function as cognitive neuroscience's goal (10). Similarly, another respondent said that in 50 years or so, we will have very precise descriptions of how the mind works (11). Interestingly, this same respondent also highly valued simple models and stated that the degree of complexity should depend on the questions you want to answer.

One respondent valued complexity in models as the world and the phenomena under study might require it (10). However, they also said that we understand complexity by abstraction and simplification, leaving it unclear whether more detailed models are desirable or not. Some respondents observed that complexity is a flexible concept. One aspect is the state of technology: More advanced computers can handle more complex models (9, 15). Another aspect is familiarity: The more you play with a complex model, the better you come to understand it (12).

A number of respondents expressed a dislike for complexity (3, 4, 5, 6, 13, 14). The leading sentiment here was that the purpose of modeling is to reduce the "dimensionality" of the phenomena under study; models that are too complex to understand simply defeat their purpose, even if we have computers to handle them (6).

A number of respondents tended toward a neutral position on complexity: How much complexity is required, or can be tolerated, depends on the goal of the model and on the balance with other virtues (1, 9, 7, 11, 12, 15).

Thirteen out of 15 respondents expressed their liking for mechanistic explanations (1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15). This relates both to causal detail and biological plausibility: According to many of the respondents, causal explanations of behavior are ultimately to be found at the level of biology.

In summary, there is no convergence of opinion regarding causal detail. Regarding the complexity that comes with causal detail, some respondents are opposed to it,

whereas others are open to it. Respondents agree that the required level of detail depends on the modeler's research question. Regarding mechanistic explanation, respondents were almost unanimously in favor of it.

The mixed response suggests that researchers want mechanistic explanations to break a phenomenon down into its underlying processes, but without going into details. Researchers do not like black boxes, but to get rid of these, it is enough to sketch the outlines of a mechanism in terms of abstract, simplified biological principles. How the balance between understanding and causal detail should be struck will depend on the goal that the model is intended to serve.

Biological Plausibility in the Interviews

Eleven out of 15 respondents value biological plausibility (respondents 1, 2, 4, 5, 7, 8, 9, 11, 12, 14, and 15), with five of these stating it is a necessary condition for any model (4, 5, 8, 11, and 14). Those valuing it often add that models should be "abstractly consistent" with biology: Models are not required to take in biophysical details, but they certainly should not violate these. One respondent explained that this is important for constraining the range of parameter values in computational modeling (1).

Two respondents stated that biological plausibility is not required of a model; a cognitive-level, functional account can be perfectly satisfactory (3, 13). Six respondents, in contrast, stated that biological plausibility is essential for mechanistic explanation, which they said is the only true form of explanation (1, 5, 8, 11, 12, 14).

Some respondents observed that biological plausibility is context dependent and often serves as a stick to beat others with (2, 5, 8, 9, 15). What you take to be the relevant level of biological detail depends on the phenomenon you are interested in; researchers interested in different aspects of the phenomenon may criticize you either for including or for excluding biological detail.

Summarizing, biological plausibility appears to be a highly valued virtue among the respondents, but mainly in the sense that models should be consistent with our knowledge of how the brain works. This tallies with the discussion of causal detail: To unpack black boxes in a computational model, it suffices to know whether the required mechanisms could be neurally implemented in principle.

DISCUSSION

To stimulate discussion on models' qualities, and thereby to enhance effective comparisons and valuation of models, we presented six key virtues in computational modeling: empirical accuracy, empirical scope, functional analysis, causal detail, biological plausibility, and psychological plausibility. After defining the virtues, we illustrated their use in Stroop modeling, where virtues are typically mentioned to provide ad hoc support for favoring one particular model over another. There is no systematic reflection on why one particular virtue of a model is taken into consideration rather than another, which virtues computational models should have across the board, or which trade-offs between virtues are to be expected.

We then presented the results of interviews with 15 experts in computational modeling, revealing that modelers give much more thought to the value and interrelations of virtues than is reflected in the literature. They are aware that the value of each virtue depends on the particular research question a model is intended to address, as well as on its relations to other virtues: Optimizing for one virtue is likely to diminish another. We briefly review some examples here.

One example of trade-off considerations is the status of empirical accuracy. Although prominent in published work, we found that empirical accuracy does not get highest priority. Researchers value a good match with the data, but certainly not at all costs. Empirical scope and interpretability often take precedence over empirical accuracy. Moreover, researchers are aware that there is a big gap between fitting the data and explaining them (cf. Chirimuuta, 2021; Thompson, 2021; Bennett, Silverstein, & Niv, 2019).

A desire to optimize for causal detail often sparks additions to unpack the black boxes in existing models, thereby also making them more complex and less easy to understand. The researchers we interviewed were of mixed opinion here, some valuing understanding above complexity, others embracing complexity or arguing that it is a relative concept. Which of these virtues should prevail depends on the research question the model is intended to answer (e.g., Francken, Slors, & Craver, 2022; Potochnik, 2015).

Biological plausibility (closely related to causal detail in cognitive neuroscience) was highly valued by our respondents, but mainly in the sense that models should be consistent with what we currently know about the brain. Some respondents observed that biological plausibility is a relative concept and that the level of biological detail you want your model to capture depends on the questions you want to answer (for discussion, also see Love, 2021).

As pointed out above, it was surprising to find that none of the respondents mentioned the trade-off between empirical scope, functional analysis, and causal detail. Another trade-off that was not mentioned in the interviews, but that is now attracting attention in the literature, is that between generalizability and interpretability. In the context of reinforcement learning models, for example, this is discussed by Eckstein and colleagues (2022), and in more general terms by Hasson, Nastase, and Goldstein (2020) and Francken and colleagues (2022).

A final takeaway is that the choice in which virtue(s) to select for depends on the goal of the modeler. This resonates with some of the current literature (e.g., Baribault & Collins, 2023; Kording, Blohm, Schrater, & Kay, 2020; Bennett et al., 2019; Wilson & Collins, 2019).

Conclusion

The reasons for making specific modeling choices remain mostly implicit in the literature, creating a situation in which different research traditions seem to be locked into their own preference bubbles. To see if this situation can be improved, we queried established experts about their ideas on model virtues, that is, the qualities of computational models that are particularly valued. We found that the considerations that go into modeling choices are actually much richer than is typically acknowledged in the literature. We presented reflections on six key virtues to illustrate the importance of a more systematic discussion of their interrelations to facilitate communication across research traditions. Although our focus here was on the field of cognitive control, we hope that our findings will spark discussion of virtues in other fields as well. We recommend that researchers make the reasons for their modeling choices more explicit in published work.

APPENDIX A: DESIGN OF SEMISTRUCTURED INTERVIEWS

Methods

Candidates for the interviews were selected according to a number of criteria:

- has published multiple widely cited computational modeling articles as first or last author, in the field of, or related to, cognitive control;
- not from the same research group;
- as much as possible represent different universities, countries, continents, modeling traditions, and genders.

The interviews were exploratory, serving as a reference for the virtues we came across in the literature. They helped us to see whether we overlooked something, and how important the identified virtues were to a selection of modelers. As such, and given that this is a qualitative and not quantitative study, the sample size was modest. Twenty-two candidates were invited (six female participants, 27%), resulting in 14 interviews with 15 respondents (two female participants, 13%). There was one interview with two respondents simultaneously; the others were one-on-one.

Procedure

All respondents were given information on the goal of the interview, the recording and storage of data, as well as a list of possible topics. Each gave written consent to participate and to have the interview recorded. Each respondent was given the choice to have their participation be fully anonymous, or to have their contribution acknowledged in special cases, for example, when a direct quote was in order.

All interviews were held online via Zoom between April 2021 and October 2021. The interviews took 63 min, on average, with the shortest interview taking 39 min, and the longest 100 min. The interviews were semistructured and were conducted by the first author of this article. A list of predetermined questions formed the basis for each interview, along with an opening and closing statement. The resulting guideline for the interviews can be found above. On the basis of the respondent's input, the order of the questions sometimes varied, sometimes questions were skipped, and some were followed up with nonlisted questions to get more in-depth information. Virtues were not primed, but there were questions designed to make them emerge.

Interview Guideline

1. Opening statement.

- Hi, welcome, thank you for your interest in this interview, and thank you for making time in your schedule!
- I am researching considerations in computational models of cognitive control processes. I am investigating the literature, but knew that that by itself would not give me the full picture. You are among the experts with whom I will be discussing views on what is important in computational models of cognitive processes, both in terms of qualities of the model, as goals it ought to achieve.
- It is important to know that there really are no right or wrong answers. I am interested in what you think, experience, and do. I am interested in the reasons you have for thinking and doing so. I have no preconceived notion of the way it ought to be.
- I might interrupt you now and then, to ensure that we focus on the goal of the interview.
- Your answers are incredibly valuable to my research, but at any point you can decide to not answer a question, or to discontinue the interview, without any consequences for you.
- The interview will be recorded, for sake of accurately keeping track of information. You might spot me taking some notes to structure the interview along the way.
- *For those who indicate they want to be credited:* You indicated that you wish to be credited in case you mention considerations that are not naturally merged with the others. To clarify this a little more: This does not mean that necessarily, you will be personally mentioned in a publication following from the interviews. I will do my best to make a coherent, integrated story out of the interviews. Still, it could be that we discuss something here today that is relevant to the story but that nobody else has brought up, or that I would like to quote something you said. In that case, you will be credited.
- Know that the content of this interview will be treated strictly confidentially, and that all files will be securely stored. Shall we start? [RECORDING STARTS]

2. Topics to discuss & questions.

2.1 Demographic information.

- Educational background
  - What university degree or degrees do you have?
- Occupational background
  - In what field would you say are you employed?
  - Have you been employed in other fields? (which?)
- How long have you been involved in computational modeling?
- Can you briefly describe what your research is about?
- Can you describe what it is you do in doing research? What is your (current) role in an experiment/paper/research project? E.g., conceptual, modeling, overseeing?
- Can you walk me through *your* process of constructing an experiment for a certain hypothesis? Perhaps in form of a timeline:
  - When do you look into the literature (if at all)?
  - When do you come up with the design for the experiment?
  - When do you think about the computational model?

2.2 Virtues as seen by respondent.

- What (types of) models do you like to use?
  - Is there a reason you chose these / prefer these?
  - Based on literature? Education?
- Can you describe the process that leads to a final computational model [for a given dataset]?
  - How do you select a certain type of model?
  - Do you modify it along the way (and how? And why?)
  - How do you decide which parameters to tweak?
- When you are designing or applying a computational model, what do you make sure it has?
- In your opinion, what are necessary qualities in a computational model? [both practical and theoretical]
- In your opinion, what qualities are desirable in a computational model? [both practical and theoretical]
- Are there qualities you think should be implemented more but are, in practice, hard to implement?
  - What do you think are the obstacles?
- What makes a model stand out among others?
- What do we come to understand by computational models?
  - What is it that we gain understanding of?

2.3 Views on goals (in science, of models).

- What do you consider to be the goals of computational cognitive neuroscience?
  - What is the main goal?
  - What are the subgoals / day-to-day-goals
- In your opinion, what is a computational model supposed to do?
  - How should a computational model attain that goal?
- What is explanation to you?
- Can you reflect on the value of a model or theory is strictly predictive?
  - Does it have a role in attaining your goals?
- Can you reflect on the value of a model or theory that is strictly descriptive?
  - Does it have a role in attaining your goals?
- In your opinion, how important is it for a model or theory to provide an explanation? And what kind of explanation?

2.4 Philosophy of science.

- What is the relation between model and theory?
- Should modelers aim to converge to one model, or can various models continue to co-exist?
- Can you reflect on the relationship between a computational model, behavioral data, and neural findings?
  - Does the model provide the mapping between behavioral and neural data?
- What is the relationship between a model and the real world?
  - Are models tools for understanding, but not to be interpreted realistically, are they actual representations or instantiations of cognitive or neural processes?

2.5 Cool down question.

- What work inspires you in computational modeling?

3. Conclusion.

- Do you have any interview-related questions for me?
- Thanking the respondent for their time and energy

Sample Characteristics

The respondents had between 10 and 40 years of experience in computational modeling, with a mean of 24.7 years. Out of the 15 respondents, two were female. Models and architectures used by the respondents were: ACT-R, connectionist or (deep) neural network models, evidence accumulation models, reinforcement learning models, statistical machine learning models, (hierarchical) Bayesian models, dynamical causal models, symbol processing models, and production system models.

The respondents had varying educational backgrounds: biology, biomedicine, cognitive neuroscience, cognitive

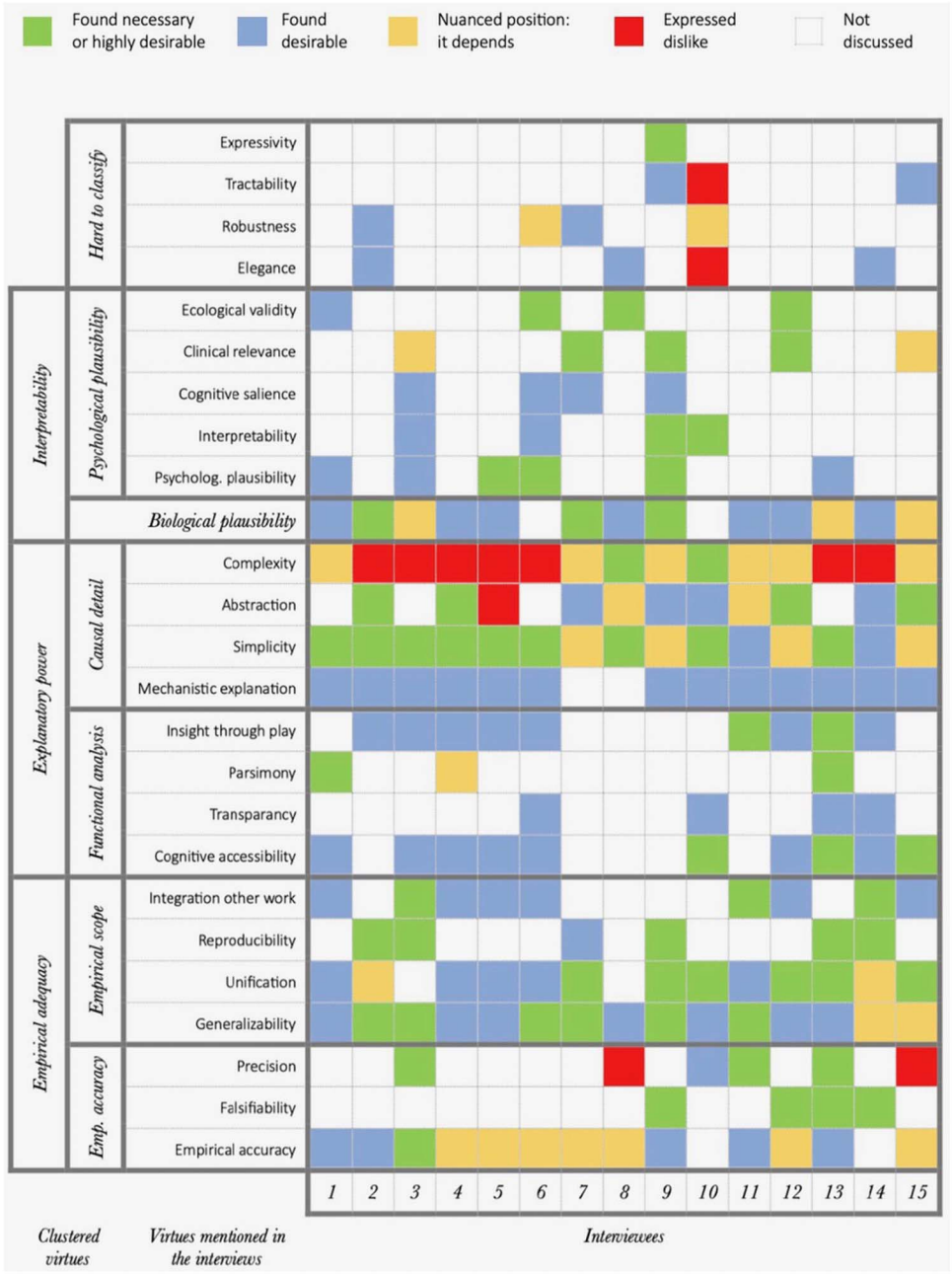
psychology, cognitive science, computational neuroscience, computer science, electrical engineering, mathematical psychology, mathematics, medicine, neuroscience, philosophy, physics, and psychology. Some studied two or more disciplines.

Self-described current fields also varied (some identified with multiple): artificial intelligence, cognitive neuroscience, cognitive psychology, cognitive science, computational cognitive neuroscience, computational cognitive psychology, computational neuroscience, computer science, decision neuroscience, neuroscience, and psychology.

Data Processing

Initial transcription was done by speech-to-text algorithm (AMBERscript). The transcriptions were edited to correct for mistakes. Before the interviews, a list of possible virtues was created based on empirical and philosophical literature, and a definition was established for each. With this as a lens, the transcripts were then labeled using ATLAS.ti. For virtues that came up that were not on the list, a definition was added to the list. After this, the data were integrated in a topic-based manner. Because of possible identifiability of respondents based on the transcripts, the data will not be made publicly available.

APPENDIX B: OVERVIEW OF VIRTUES DISCUSSED IN THE INTERVIEWS



Appendix B. Overview of virtues discussed in the interviews.

Corresponding author: Saskia Heijnen, Leiden University Faculty of Social and Behavioural Sciences, Wassenaarseweg 52, Leiden, Zuid-Holland, Netherlands, 2333 AK, or via e-mail: s.heijnen@outlook.com.

Data Availability Statement

Interview data gathered for this study will not be made available, because of possible identifiability of the respondents.

Author Contributions

Saskia Heijnen: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing—Original draft; Writing—Review & editing. Jan Sleutels: Conceptualization; Methodology; Supervision; Visualization; Writing—Review & editing. Roy de Kleijn: Conceptualization; Methodology; Supervision; Visualization; Writing—Review & editing.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience (JoCN)* during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be: M/M = .606; W/M = .061; M/W = .091; W/W = .242.

REFERENCES

Baribault, B., & Collins, A. G. E. (2023). Troubleshooting Bayesian cognitive models. *Psychological Methods*. https://doi.org/10.1037/met0000554, PubMed: 36972080

Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The two cultures of computational psychiatry. *JAMA Psychiatry*, *76*, 563–564. https://doi.org/10.1001/jamapsychiatry.2019.0231, PubMed: 31017638

Blohm, G., Kording, K. P., & Schrater, P. R. (2020). A how-to-model guide for neuroscience. *eNeuro*, *7*, ENEURO.0352-19.2019. https://doi.org/10.1523/ENEURO.0352-19.2019, PubMed: 32046973

Bokulich, A. (2017). Models and explanation. In L. Magnani & T. Bertolotti, *Springer handbook of model-based science* (pp. 103–118). Springer. https://doi.org/10.1007/978-3-319-30526-4_4

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652. https://doi.org/10.1037/0033-295X.108.3.624, PubMed: 11488380

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*, 767–790. https://doi.org/10.1007/s11229-020-02713-0

Chuderski, A., & Smolen, T. (2016). An integrated utility-based model of conflict evaluation and resolution in the Stroop task. *Psychological Review*, *123*, 255, 290. https://doi.org/10.1037/a0039979, PubMed: 26751852

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332–361. https://doi.org/10.1037/0033-295X.97.3.332, PubMed: 2200075

Collins, A. G. E., & Shenhav, A. (2022). Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, *47*, 104–118. https://doi.org/10.1038/s41386-021-01126-y, PubMed: 34453117

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2022). The interpretation of computational model parameters depends on the context. *eLife*, *11*, e75474. https://doi.org/10.7554/eLife.75474, PubMed: 36331872

Fennell, A., & Ratcliff, R. (2019). Does response modality influence conflict? Modelling vocal and manual response Stroop interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 2098–2119. https://doi.org/10.1037/xlm0000689, PubMed: 30802093

Francken, J. C., Slors, M., & Craver, C. F. (2022). Cognitive ontology and the search for neural mechanisms: Three foundational problems. *Synthese*, *200*, 378. https://doi.org/10.1007/s11229-022-03701-2

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*, 416–434. https://doi.org/10.1016/j.neuron.2019.12.002, PubMed: 32027833

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*, 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Jiang, J., Heller, K., & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neuroscience & Biobehavioral Reviews*, *46*, 30–43. https://doi.org/10.1016/j.neubiorev.2014.06.001, PubMed: 24929218

Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, *195*, 2761–2793. https://doi.org/10.1007/s11229-017-1355-6

Kording, K. P., Blohm, G., Schrater, P., & Kay, K. (2020). Appreciating the variety of goals in computational neuroscience. *arXiv*. https://doi.org/10.48550/arXiv.2002.03211

Lawler, I., & Sullivan, E. (2021). Model explanation versus model-induced explanation. *Foundations of Science*, *26*, 1049–1074. https://doi.org/10.1007/s10699-020-09649-1

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*, e1006043. https://doi.org/10.1371/journal.pcbi.1006043, PubMed: 29694347

Love, B. C. (2021). Levels of biological plausibility. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *376*, 20190632. https://doi.org/10.1098/rstb.2019.0632, PubMed: 33190602

Lovett, M. C. (2002). Modeling selective attention: Not just another model of Stroop (NJAMOS). *Cognitive Systems Research*, *3*, 67–76. https://doi.org/10.1016/S1389-0417(01)00045-6

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*, 975–995. https://doi.org/10.1007/s11229-011-0054-y

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203. https://doi.org/10.1037/0033-2909.109.2.163, PubMed: 2034749

Marr, D. C. (1982). *Vision*. San Francisco: Freeman.

Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, *136*, 107261. https://doi.org/10.1016/j.neuropsychologia.2019.107261, PubMed: 31733237

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

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, *135*, 601–609. https://doi.org/10.1037/bne0000471, PubMed: 34096743

Niv, Y., & Langdon, A. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67–73. https://doi.org/10.1016/j.cobeha.2016.04.005, PubMed: 27408906

Peebles, D., & Cooper, R. P. (2015). Thirty years after Marr's Vision: Levels of analysis in cognitive science. *Topics in Cognitive Science*, *7*, 187–190. https://doi.org/10.1111/tops.12137, PubMed: 25755213

Poggio, T. (2010). Afterword: Marr's vision and computational neuroscience. In D. Marr (Ed.), *Vision* (pp. 362–365). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262514620.003.0009

Potochnik, A. (2015). The diverse aims of science. *Studies in History and Philosophy of Science*, *53*, 71–80. https://doi.org/10.1016/j.shpsa.2015.05.008, PubMed: 26386532

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/2004.001.0001

Rosales, A., & Morton, A. (2021). Scientific explanation and trade-offs between explanatory virtues. *Foundations of Science*, *26*, 1075–1087. https://doi.org/10.1007/s10699-019-09645-0

Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press. https://doi.org/10.1017/9781108381352

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*, 217–240. https://doi.org/10.1016/j.neuron.2013.07.007, PubMed: 23889930

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. https://doi.org/10.1037/h0054651

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

Thompson, J. A. F. (2021). Forms of explanation and understanding for neuroscience and artificial intelligence. *Journal of Neurophysiology*, *126*, 1860–1874. https://doi.org/10.1152/jn.00195.2021, PubMed: 34644128

Verguts, T. (2022). *Introduction to modeling cognitive processes*. Cambridge, MA: MIT Press.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547, PubMed: 31769410

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*, 201–219. https://doi.org/10.1007/s11098-008-9324-z