# Enhancing Methodological Rigor for Computational Cognitive Science: Core Tenets and Ad Hoc Residuals

**Jennifer Roberts (jenmarie@mit.edu)**
MIT CSAIL, 32 Vassar Street
Cambridge, MA 02139 USA

**Jacob Beal (jakebeal@bbn.com)**
BBN Technologies, 10 Moulton Street
Cambridge, MA 02138 USA

## Abstract

Computational models are notoriously difficult to compare and interpret, resulting in a community segmented around modeling paradigms. In this paper, we seek to develop community standards and methodology that will make it easier to compare work across computational paradigms, discern what types of empirical predictions can be drawn from computational work, and test the validity of computational models. Using an established Bayesian model, we illustrate how our proposed methods will achieve these goals.

**Keywords:** Methodology; Marr Levels; Core Tenets

## Introduction

As cognitive scientists, we seek to develop a unified theory of the human mind. Our current modeling efforts, however, use a diverse set of tools and formalisms that span from neural networks to cognitive architectures to Bayesian reasoning to first order predicate logic, and each model only addresses aspects of the larger problem. In this paper, we seek to develop community standards that will make it easier to compare work across paradigms, discern what types of empirical predictions can be drawn from computational work, and enhance our ability to translate models into statements about the computational nature of intelligence.

In the first section, we introduce two new terms: *core tenet*, for a part of a computational model argued to be cognitively plausible, and *ad hoc residual*, for an arbitrary implementation detail. We then describe two types of core tenets: *central assumptions* that an author believes to be paramount to the cognitive theory behind their computational model and *peripheral hypotheses* that the author believes are cognitively plausible but feels less committed to preserving in their current form. We advocate the explicit delineation of a model's core tenets and ad hoc residuals as well as the explicit division of core tenets into central assumptions and peripheral hypotheses when publishing computational papers, because we believe this practice will enable computational researchers to communicate more effectively and to learn more from one another's work. Identification of the intended core tenets and ad hoc residuals will reduce the chance that other researchers will take issue with an aspect of a model that the authors consider tangential to their work. We argue that this practice will allow researchers to clearly communicate the correspondence between their models and theories of human cognition, establish the validity of one another's theories by using multiple modeling formalisms to instantiate the same theory, and establish whether a computational model's performance is truly independent of arbitrary design decisions.

## Cognitive Plausibility

Computational models must be defined at a finer-grained level of detail than other types of psychological models, because computational models must lead to software implementations or sets of mathematical computations. Defining such a model almost always entails specifying details for which the modeler has little or no psychological or neurological evidence (Newell, 1990; Anderson, 1983; R. Cooper, Fox, Farringdon, & Shallice, 1996; McClelland, 2009).

Because details for which we have no evidence must be specified in order for the model to be implemented, computational models contain both cognitively plausible elements and ad hoc implementation details. We call the parts of a computational model that the modeler considers cognitively plausible and wishes to communicate to the modeling community the *core tenets* of the model, and we refer to the incidental design decisions as the *ad hoc residuals*.

We further refine this idea using Lakatos' (1970) observation that theoretical claims can be broken into two groups based on the strength of a researcher's commitment to the claims. Core tenets can thus be subdivided into central assumptions and peripheral hypotheses (following the example of Cooper and Shallice (2000)). Tenets that the researcher is strongly committed to preserving, because he or she believes those aspects of the model generate crucial behavior, are called central assumptions. Tenets that the researcher regards as open to modification are called peripheral hypotheses. While an author writing about a model may have a clear idea of which parts they intend as central or peripheral core tenets and which parts they intend as ad hoc residuals, that intent often is not communicated clearly.

Cooper (2006, 2007) highlights the difficulty of identifying core tenets in another person's work. Cooper analyzes the progression of the cognitive theories behind Soar and ACT-R, labeling architectural aspects that remained stable over time as central assumptions and aspects that were added to accommodate empirical findings as peripheral hypotheses. While this is a reasonable way to proceed when the original author has not explicitly delineated core tenets, Cooper's criteria for identifying central assumptions confounds the ease of mod-

| Core Tenets | Ad Hoc Residuals |
|---|---|
| Cognitively plausible parts of the computational model, which specify a testable cognitive theory. The core tenets should be verified both empirically and computationally. As a cognitive theory, core tenets can be divided into central assumptions and peripheral hypotheses. | Implementation details that are based on minimal cognitive or neurological evidence and are not intended to be part of a cognitive theory. The success of the computational model should be shown to be independent of the ad hoc residuals. |

| Central Assumptions | Peripheral Hypotheses |
|---|---|
| Aspects of the cognitive theory that the scientist will only change as a last resort | Aspects of the cognitive theory that the scientist will update based on empirical findings |

Figure 1: Framework through which any computational model of cognition may be understood. This framework distinguishes the parts of a computational model that the modeler considers cognitively plausible from the parts he or she deems cognitively-irrelevant. Because core tenets specify a theory of cognition, the core tenets can be divided into central assumptions and peripheral hypotheses based on a scientist's level of commitment to preserving the tenets in their current form (Lakatos, 1970).

ifying computational models with the question of cognitive plausibility.

Aspects of a computational model might remain stable for engineering reasons, so the stability of a particular aspect is not a reliable metric for ascertaining whether the modeler considers the aspect cognitively plausible. For example, pre-2004 versions of Soar relied exclusively on production rules. Does this imply that production rules were viewed as the only representation used by the brain, a reasonable approximation for the brain's symbolic representations, or an engineering tool that could be modified or replaced if necessary? Citing the continual use of production rules as evidence for its status as a core tenet ignores the possibility that first order logic was initially adopted and continues to be used for engineering simplicity. Cooper's efforts to identify the central assumptions and peripheral hypotheses of Soar highlight the need for researchers to unambiguously identify their core tenets, because despite Cooper's objective approach, identifying which parts of a computational model are intended to be cognitively plausible still involves guess work.

Throughout the remainder of this paper, we focus on the distinction between core tenets and ad hoc residuals, that is, the divide between the cognitively plausible parts of a model and the details added for implementation purposes. We agree with Cooper that further dividing core tenets into central assumptions and peripheral hypotheses is extremely important and advocate using this terminology when defining models. In later sections, we will focus on determining the extent to which either type of core tenet affects a model's performance and methods for comparing either type of core tenet across modeling paradigms.

Marr partially addressed the divide between core tenets and ad hoc residuals when he proposed that computational models can be analyzed using the following three levels: (1) the computational theory, which includes the goal of the computation and the general strategy for performing the computation, (2) the representation and algorithm, which includes the input and output representations as well as the algorithm for manipulating those representations, and (3) the hardware implementation, which includes the physical realizations of the representations and algorithms (Marr, 1982). He argued that some modelers work at the hardware level while others work at the algorithmic or computational level. This implies that a modeler working at the computational theory level would consider the goal of their algorithm to be a core tenet and would consider the algorithmic and hardware details to be ad hoc residuals. In contrast, a modeler working at the algorithmic level would have core tenets that involve representations and algorithmic details and ad hoc residuals related to the hardware implementation.

While a useful construct, the Marr levels still leave room for ambiguity. Every researcher can have a slightly different interpretation of where the separations between levels belong. For example, Broadbent (1985) and Rumelhart and McClelland (1985) debated about which Marr levels existing cognitive theories and connectionist models really addressed.

Further complicating the issue, complex models contain processes with multiple subgoals, so the overall process and each of the subprocesses all have Marr computational level descriptions. For example, the Companion cognitive architecture has been used to perform transfer learning (Forbus & Hinrichs, 2006). At the computational level, the overall system attempts to solve problems by analogically comparing problems to previously-encountered examples. The system contains an analogical component with its own computational level, which focuses on comparing structural elements to determine the similarity between two problems. Thus, the Companion architecture might suggest a computational level theory, but are all the computational level details relevant to the cognitive theory, or only a subset of them?

Claiming that two models pertain to distinct Marr levels implies that the models have different core tenets and ad hoc residuals, but the Marr level descriptions do not unambiguously specify what is different (a point also noted in McClelland (2009)). Currently, the phrase "modeling at the computational level" merely indicates that some details are considered ad hoc residuals without explicitly explaining which details fall into this category.

## An Illustrative Cognitive Model

In the remaining sections, we will show how core tenets and ad hoc residuals can (1) clarify assertions about cognitive plausibility, (2) provide mechanisms for testing how much of a computational model's success depends on its cognitively plausible parts and how much of its success depends on incidental design decisions, and (3) specify empirically-testable cognitive theories. To illustrate our points, we use a relatively straight-forward computational model of how people play a simple number game (Tenenbaum, 2000). This analysis applies equally well to any computational model of cognition, regardless of its paradigm or degree of complexity.

Tenenbaum (2000) explores how people play a game in which they receive sets of numbers between 1 and 100, like {4, 8, 24, 12}, and guess what rule generated the set. Possible rules include "multiples of 3," "squares," and "numbers between 10 and 20." In this case, the numbers satisfy several rules, including "multiples of two," "multiples of four," and "numbers less than 30." To ascertain how people think the set has been generated, each person receives a series of test numbers, like {3}, {22}, and {16}, and they must rate the likelihood that each test number belongs to the set.

Tenenbaum proposed a computational model in which people start with a set of hypotheses about how sets have been generated. Hypotheses come in two forms, those that depend on a mathematical rule and those that involve an interval. The prior probability assigned to the rule-based set is uniformly distributed across all rule-based hypotheses, while the weight assigned to the interval-based hypotheses is distributed so that intermediate-sized intervals receive the bulk of the prior probability.

The probability of observing a set of numbers, $X$, given that a particular hypothesis generates the set, depends on how many numbers between 1 and 100 the hypothesis can generate. For example, the hypothesis "multiples of 2" can generate 50 numbers between 1 and 100, while the hypothesis "numbers from 1 to 10" generates 10. High likelihoods are assigned to hypotheses that fit the observed data without generating too many extra numbers. Thus, {10, 30, 70} would be more likely to be generated by "multiples of 10" than by "multiples of 2." This model builds on what Tenenbaum refers to as the size principle, a trade-off between simplicity and goodness of fit, which essentially states that people prefer the simplest hypothesis that explains the data.

## Using Core Tenets to Indicate Intent

Just by examining his computational model, we do not know which parts of the model Tenenbaum considers cognitively plausible. By explicitly stating a set of core tenets, however, we can indicate which parts of a model readers should focus on when evaluating a piece of computational work.

For example, if we take the trade-off between simplicity and goodness of fit as the only core tenet, all other aspects of the model would be considered ad hoc residuals. Under this interpretation, the Bayesian framework would be a useful language for describing the model, but any other framework in which we could describe a similar trade-off between simplicity and goodness of fit would work equally well. Readers assessing the cognitive plausibility of the work should then evaluate the extent to which they believe a trade-off between simplicity and goodness of fit adequately captures a crucial component of how people play the number game, without allowing the Bayesian framework or particular probability distributions to play a crucial role in their evaluation process.

Realistically, most computational models involve a set of core tenets. Based on Tenenbaum's description of his model, we can identify a few principles that seem like compelling core tenet candidates, such as the trade-off between simplicity and goodness of fit (described in the paper as the size principle) and the ability to process both similarity-based and rule-based hypotheses using the same apparatus. Whether or not other aspects of the model or Bayesian framework should be considered core tenets remains a source of speculation, but explicitly specifying a list of core tenets and ad hoc residuals and indicating a level of commitment to the items in that list by dividing the core tenets into central assumptions and peripheral hypotheses would clarify Tenenbaum's perspective.

A clear division between core tenets and ad hoc residuals would also elucidate the simplifying assumptions necessary to make a modeling effort tractable. McClelland (2009) writes eloquently about the need for simplifications in order to develop a model that can test the consequences of a set of core ideas in an comprehensible manner. Labeling simplifications as ad hoc residuals will make it clear to the reader what falls within and outside of the scope of the model and what should be considered when evaluating the model.

Anyone can effortlessly spring to a controversial conclusion about a computational model. For example, do connectionist models imply that the brain uses a back propagation algorithm? Do Bayesian models imply that the brain contains a probabilistic engine? Do symbolic models imply that the brain stores information using first order logic predicates? Even though these conclusions may not reflect a modeler's beliefs, these types of reactions can lead other researchers to doubt the premise of a particular piece of work, hassle a modeler with questions that are tangential to the intent of the work, or even reject an entire modeling paradigm.

Clearly specifying the core tenets of a model can focus the ensuing conversation on the parts of a model that the authors deem most relevant and most cognitively plausible. Debates over controversial claims may be the hallmark of science, but such debates should focus on claims that the modelers decisively make, not claims that readers infer from the ad hoc residuals of a model or modeling paradigm.

## Using Core Tenets to Validate Models

Ideally, we would like to know that a model's success relies only on the parts we believe to be important and not on arbitrary design decisions, but we as a community currently lack a systematic process for making this distinction. Cooper

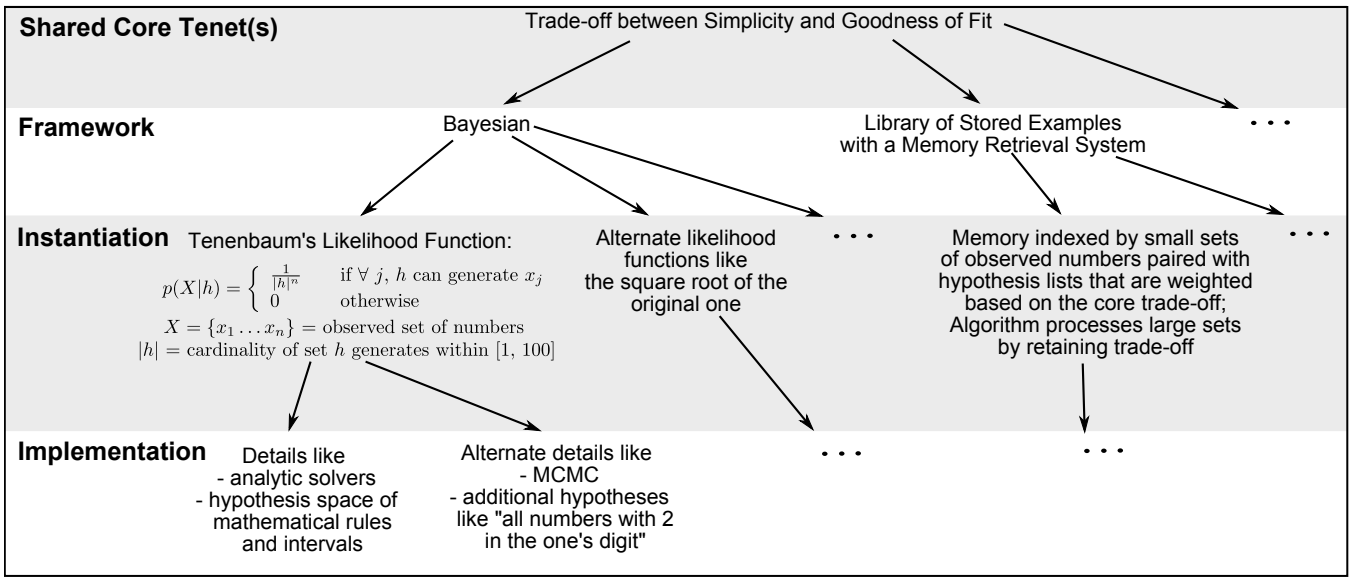


Figure 2: A shared-core-tenet analysis can be used to explore how the behavior of any computational model, regardless of its paradigm or complexity, depends on a set of core tenets. This example depicts a comparison between different formulations of the Tenenbaum (2000) number game model, where the leftmost branch describes the original formulation of the model and the remaining branches specify hypothetical alternative formulations. Other alternate formulations, like those described in Shi, Feldman, and Griffiths (2008), would be depicted using additional branches.

et al (1996) suggests using a criticality/sensitivity ratio that measures the degree to which a computational model's behavior depends on core tenets versus ad hoc residuals, but we do not see a systematic way of defining such a ratio because a model's dependence on underlying parameters and constraints is too complex to be summarized by a single ratio. In addition, Cooper's work only seeks to compare theories within a single research program, but we are interested in what we can learn by comparing research programs that share one or more core tenets but rely on different modeling paradigms. For instance, if ACT-R and Soar both share a core tenet, can that provide evidence that the shared tenet yields some desirable behavior? Similarly, if Bayesian and connectionist models for the same phenomenon share a core tenet, what does that tell us about the tenet?

In subsequent sections, we introduce a framework for systematically comparing *core-tenet-consistent models*, models that share a single core tenet or a set of core tenets, regardless of whether the models use the same paradigm. This involves developing alternate formulations of the original computational model that preserve the core tenets but change the ad hoc residuals. If the core tenets are valid, these core-tenet-consistent versions of the model should produce results that are qualitatively similar to both empirical findings and the results obtained using the original model, while providing evidence that the results do not depend on ad hoc residuals.

## Creating Core-Tenet-Consistent Models

We can develop an array of core-tenet-consistent models by first specifying a set of core tenets that will be shared by the models. As a simple example, we consider the core tenet trade-off between simplicity and goodness of fit in Tenenbaum's number game model.

We can explore the space of models consistent with the shared core tenets by using the following levels of abstraction. At the highest level, we start with a framework capable of instantiating the shared core tenets. Tenenbaum's model uses a Bayesian framework, but another formulation of the model might use a cognitive architecture or logical model to express the principles of simplicity and goodness of fit.

The instantiation level specifies how the shared core tenets are expressed within the formalism selected at the framework level. At this level of detail, Tenenbaum's model depends on the specific hypothesis space, prior probabilities, and likelihood distributions described in the previous section. Alternative models might use an expanded hypothesis space or different set of probability distributions that still preserve the core tenet trade-off.

The bottom level specifies the remainder of the implementation details necessary to turn the specific instantiation of the shared core tenets into a working computational model. Tenenbaum's model appears to compute analytical solutions for the posterior distributions, but an alternative formulation might rely on a non-deterministic algorithm like MCMC.

Also at this level, Tenenbaum's model calculates a posterior probability for all hypotheses when determining the probability that a novel number belongs to a particular set, whereas an alternative algorithm might use heuristics to focus on a subset of hypotheses.

The number of shared core tenets combined with their specificity will determine how significantly a core-tenet-consistent formulation of a model may diverge from the original version. For example, if the core tenets mirror the constraints imposed by the Bayesian framework, then all core-tenet-consistent models will either use the Bayesian framework or a pseudo-Bayesian one. In contrast, when core tenets provide looser constraints, alternative formulations will involve a wider variety of frameworks and paradigms.

### Testing the Validity of Core Tenets and the Insignificance of Ad Hoc Residuals

Figure 2 shows a sampling of the models consistent with a core tenet trade-off between simplicity and goodness of fit. If this core tenet is valid, these core-tenet-consistent versions of the model should produce results that are qualitatively similar to both empirical findings and the results obtained using Tenenbaum's original model.

Developing multiple formulations explores whether the model's performance relies on any of the ad hoc residuals. If an ad hoc residual contributes to the model's success, we would expect the residual to play a crucial part in every implementation. For example, when we presume that the trade-off between simplicity and goodness of fit is the only core tenet, this implies that everything else, including the specification of a fixed hypothesis space, is an ad hoc residual. If the specification of a fixed hypothesis space is in fact crucial to Tenenbaum's model, every core-tenet-consistent model, defined with respect to simplicity and goodness of fit, should contain some type of fixed hypothesis space. In this case, even if a fixed hypothesis space had initially been considered an ad hoc residual, its presence in every core-tenet-consistent model might imply that it should be considered a core tenet. However, if an ad hoc residual that does not seem cognitively plausible still plays a crucial part in every instantiation, we would need to search for cognitively plausible alternatives.

## Core Tenets Specify Cognitive Theories

Ultimately, we would like to develop cognitive theories that describe an individual's cognitive processes and make useful empirical predictions, but what types of predictions can legitimately be drawn from a detailed, fully-implemented computational model? How can we systematically reconcile computational work with non-computational work?

Core tenets identify which parts of a computational model describe an individual's cognitive processes. For example, Tenenbaum's paper does not directly indicate how his model relates to individual cognition (Tenenbaum, 2000), but Figure 3 shows two possible interpretations, specified by two expanded sets of core tenets, both of which translate into cognitive theories of how individuals play the number game.

To verify the bottom set of core tenets, one must demonstrate that a Bayesian model can produce results that match the performance of each individual participant of the number game experiment. Tenenbaum's original formulation compared results obtained using his Bayesian model to results obtained by averaging over a population. This suggests that his model captures a population average, but it may or may not adequately capture what any one person is doing. To establish that these core tenets describe an individual's cognitive process, one would have to establish that a particular formulation consistent with the core tenets can be used to adequately model each individual's performance. The conglomerate set of formulations, each tuned to a specific person's performance, should then provide a computational account for the group performance. Thus, discerning whether the bottom set of core tenets provides a good characterization of individual behavior or simply a description of aggregate group behavior would probably require additional empirical tests.

### Testing Computational Plausibility Claims

When designing a complex computational model, one often starts with a set of implicit assumptions, perhaps philosophically based, that constrain the design of the system. When a system fails to perform as anticipated, these implicit assumptions may change or become more constrained. Computational work of this type seeks to answer the question: Can we accomplish task $X$ under the set of constraints $Y$?

Depending on task complexity and the initial sets of constraints, surprising failures may yield insights into the computational nature of intelligence, but publishing negative results remains difficult. Core tenets provide a systematic language for describing this type of work. By contrasting the core tenet constraints necessary for building a working system with the sets of constraints for which no working system can be found, we can begin to systematically explore the space of constraints required for functional cognitive models.

## Contributions

We have introduced the terms core tenets and ad hoc residuals to distinguish cognitively plausible parts of a computational model from incidental implementation details, and we have demonstrated how these concepts can augment and validate work on a well-known computational model (Tenenbaum, 2000).

We argue that computational cognitive scientists should explicitly identify core tenets and ad hoc residuals and distinguish between central and peripheral core tenets when describing their models. This practice will elucidate an author's intent, provide mechanisms for systematically testing whether the core parts of a computational model play an instrumental role in the model's success, and help ensure that a model's performance remains independent of arbitrary implementation details. Using core tenets and ad hoc residuals can (1) bridge between modeling paradigms by helping researchers to create core-tenet-consistent instantiations of a model, (2) bridge between computational and

| Sample Set of Core Tenets | Cognitive Theory Implied by the Tenets |
|---|---|
| 1. Specification of a hypothesis space for each set of numbers encountered during the game<br>2. Methods for weighting hypotheses based on a trade-off between simplicity and goodness of fit and for favoring intermediately-sized intervals over small or large intervals | Each person has a method for defining a personal hypothesis space, and people may use different hypothesis spaces depending on context. Given a set of observed numbers, people weight the likelihood of each hypothesis using a trade-off between simplicity and goodness of fit, as well as a tendency to favor intermediately sized intervals over small and large intervals. For a particular set of numbers, people may only consider a subset of the hypotheses they use throughout the course of the game. |
| 1. A unique pre-defined hypothesis space for each individual that remains fixed throughout the game<br>2. A set of prior and likelihood functions defined over the hypothesis space in a manner that favors simplicity, goodness of fit, and intermediate interval size<br>3. An algorithm for considering the posterior probability of *all* hypotheses | Each individual has a personal hypothesis space, and people reuse the same space every time they play the number game. Each individual uses a personal prior and likelihood function to weight the space, and all individuals use the same posterior probability equation for comparing hypotheses. For every set of numbers, each person considers every hypothesis in his or her hypothesis space, regardless of how large the space is. |

Figure 3: Core tenets specify theories about an individual's cognitive processes. The sets of core tenets in this figure describe two possible interpretations of the Tenenbaum (2000) model, which translate into two theories of how an individual plays the number game. Because Tenenbaum's model instantiates both sets of core tenets, his model is consistent with both theories.

non-computational work by clearly indicating how a computational model translates into a theory of individual cognition from which we can draw empirical predictions, (3) avoid misconceptions by focusing scientific debate on claims we intentionally make instead of claims that appear to be implied by our models, and (4) provide a theoretical framework for testing claims about computational plausibility with respect to a set of constraints.

We expect that providing clear statements about the core tenets and ad hoc residuals of our models will greatly enhance the ability of cognitive scientists to communicate and to compare work across paradigms, and we strongly encourage the community to adopt and enforce this standard.

## Acknowledgements

The authors thank Richard Cooper, Josh Tenenbaum, and three anonymous reviewers for their helpful comments. J.M.R thanks the Fannie and John Hertz Foundation and both authors thank the National Science Foundation for their financial support.

## References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General*, *114*(2), 189-192.

Cooper, R., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence*, *85*, 3-44.

Cooper, R. P. (2006). Cognitive architectures as Lakatosian research programs: Two case studies. *Philosophical Psychology*, *19*, 199-220.

Cooper, R. P. (2007). The role of falsification in the development of cognitive architectures: Insights from a Lakatosian analysis. *Cognitive Science*, *31*, 509-533.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 297338.

Forbus, K. D., & Hinrichs, T. R. (2006). Companion cognitive systems: A step toward human-level AI. *AI Magazine*, *27*(2).

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (p. 91-196). Cambridge, UK: Cambridge University Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt and Co., Inc.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*, 11-38.

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

Rumelhart, D. E., & McClelland, J. L. (1985). Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General*, *114*(2), 193-197.

Shi, L., Feldman, N. H., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In *30th Annual Conference of the Cognitive Science Society.*

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K. R. Muller (Eds.), *Advances in Neural Information Processing Systems 12* (pp. 59–65). Cambridge, MA: MIT Press.