TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 7 (2015) 230–242

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12131

The Algorithmic Level Is the Bridge Between Computation and Brain

Bradley C. Love

Experimental Psychology, University College London

Received 18 August 2013; received in revised form 27 April 2014; accepted 18 June 2014

Abstract

Every scientist chooses a preferred level of analysis and this choice shapes the research program, even determining what counts as evidence. This contribution revisits Marr's (1982) three levels of analysis (implementation, algorithmic, and computational) and evaluates the prospect of making progress at each individual level. After reviewing limitations of theorizing within a level, two strategies for integration across levels are considered. One is top–down in that it attempts to build a bridge from the computational to algorithmic level. Limitations of this approach include insufficient theoretical constraint at the computation level to provide a foundation for integration, and that people are suboptimal for reasons other than capacity limitations. Instead, an inside-out approach is forwarded in which all three levels of analysis are integrated via the algorithmic level. This approach maximally leverages mutual data constraints at all levels. For example, algorithmic models can be used to interpret brain imaging data, and brain imaging data can be used to select among competing models. Examples of this approach to integration are provided. This merging of levels raises questions about the relevance of Marr's tripartite view.

Keywords: Levels of analysis; Approximately Bayesian; Model-based fMRI analysis; Categorization; Rational analysis

Scientists are chiefly in the business of doing science as opposed to philosophizing about levels of analysis. Nevertheless, every scientist chooses, perhaps implicitly, a preferred level of analysis and this choice shapes the research program. One's preferred level of analysis determines what counts as evidence for or against a theory, the big questions to answer, and the key experiments to conduct. Thus, the relative merits of different levels of analysis and how these levels relate to one another is a topic that invites consideration from both practitioners and philosophers of science.

Correspondence should be sent to Bradley C. Love, Experimental Psychology, University College London, 26 Bedford Way, Room 235, London, UK WC1H 0AP. E-mail: b.love@ucl.ac.uk

Marr's (1982) three levels of analysis have been highly influential in shaping research in cognitive science. The three levels, namely the implementation, algorithmic, and computational levels, roughly corresponding to where/what, how, and why questions, respectively. The implementation level is concerned with the physical substrate that supports behavior. Resting on the implementation level is the algorithmic level, which is concerned with the processes and representations that give rise to behavior. In cognitive science research, these two levels can be seen as defining the *mechanism* that supports behavior (Craver, 2007). As will be discussed below, studying mechanism invites both looking "down" toward how a mechanism decomposes into parts and looking "up" toward the broader context or environment in which a mechanism is situated (Bechtel, 2009).

The computational level is the most abstract of Marr's levels and is not concerned with mechanism. The nature of the computing device (i.e., implementation level) and how the computation is carried out (i.e., the algorithmic level) are irrelevant at this level of analysis. The sole concern of the computational level is the abstract problem description, which consists of detailing the input–output relationships (i.e., for this stimulus, people will make this decision). As interest in the rational Bayesian explanations has grown (Chater, Tenenbaum, & Yuille, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011), the computational level has been repurposed as the optimal input–output function given some set of assumptions (Griffiths, Vul, & Sanborn, 2012). In other words, the computational level is now routinely interpreted as not just any abstract problem description, but as optimal in some sense. I will stick with this popular characterization of the computational level as a rational Bayesian account of cognition. Note that not every Bayesian account is a rational account. For further reading on the relationship between rational analysis and Bayesian models, please see Jones and Love (2011).

For a digital computer, the implementation level is the hardware, the algorithmic level is the program, and the computational level is the program specification. Of course, the brain is not a computer in any simple sense, and it is unclear in cognitive science whether these levels neatly separate conceptually or are even theoretically useful. Indeed, this contribution argues for a research strategy that bridges all three levels of analysis. One result of this bridging is that the necessity of Marr's distinctions is brought into question.

In the remainder of this contribution, critiques of theories that are solely phrased at one level of analysis are considered (see Cooper & Peebles, 2015, for related discussion). Then, two possibilities for bridging levels of analysis are evaluated. The first proposal, the *top–down* view, begins with computational-level theories and considers how these theories could be computed by approximate Bayesian inference to explain findings at the algorithmic level (cf. Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010). In other words, these bridging theories argue that algorithmic theories are simply computational-level theories with capacity constraints. Some challenges for this approach include that people are suboptimal for reasons other than capacity limitations, and that theories are under-constrained at the computational level and therefore do not provide a reliable foundation for theoretical integration. The second proposal for bridging levels is the *inside-out* view in which algorithmic-level theories are used to

explain implementation-level findings, including brain imaging results, as well as to link to formalisms that are optimal in some sense (i.e., reside at the computational level). Considering data at the algorithmic and implementation levels provides a number of mutual constraints on theory development. Furthermore, mechanistic models that touch on aspects of both algorithm and implementation can readily be compared to a computational-level account, facilitating integration across all three of Marr's levels.

1. Theorizing at a single level

Marr's hope was that findings across levels of analysis would eventually be reconciled and integrated. In the interim, Marr's levels of analysis were intended to be somewhat distinct as different questions are addressed at each level. However, it is unclear whether theories can be successfully developed and evaluated by strictly sticking within a single level of analysis.

The necessity of bridging Marr's levels is perhaps clearest at the implementation level. It is not clear one can formulate a theory of human behavior solely in the language of brain measures. Some theoretical entity must be localized. Brain imaging, even though somewhat removed from the physical processes supporting cognition, is essentially an exercise in localization. This has led to the criticism that cognitive neuroscience is the new phrenology (Franz, 1912; Utall, 2001). Localizing mental function need not be problematic. The issue is what to localize. The value of a theory that localizes mental function lies in both the characterization of the mental process and the bridge theory that links this characterization to the brain (Love & Gureckis, 2007). Starting with an ill-specified or folk psychological theory of mental function ultimately limits the value of the overall enterprise and invites comparison with phrenology. However, localizing a well-specified theory at the algorithmic level can be fruitful (Love & Gureckis, 2007).

One may counter that by examining the brain directly, one can devise an entirely new theoretical apparatus from the ground up that does not inherit the limitations and assumptions of current theories. However, it is hard to see how one could truly develop such a new science in a theoretical vacuum. As a thought experiment, imagine a brain imaging device that has infinite spatial and temporal resolution such that it provides accurate measurements of the state of every neuron at every moment. This amazing machine alone will not answer how to best present material to students or how to treat depression. Although children in classrooms and people with depression could be imaged, it is not clear how this deluge of high-resolution data would by itself suggest an intervention. Much like the Human Genome Project, the real work would start once this mythical imaging device was available.

Turning to the algorithmic level, one common criticism is that theories at this level are somewhat arbitrary in that numerous algorithms can generate the same output (i.e., predict certain behaviors in a situation) in the same way that an infinite number of computer programs can accomplish the same task. Echoing these sentiments, Anderson (1991) stated, "All mechanistic proposals which implement the same rational prescription are the

same," and "a rational theory provides a precise characterization and justification of the behavior the mechanistic theory should achieve." These views are seconded by Chater and Oaksford (1999): "The picture that emerges from this focus on mechanistic explanation is of the cognitive system as an assortment of apparently arbitrary mechanisms, subject to equally capricious limitations, with no apparent rationale or purpose." In practice, lessons are drawn by comparing how algorithmic models address behavior and there is a clear lineage of models. Even in extreme cases where superficially different models formally converge in their predictions, key lessons are drawn from understanding these equivalences (e.g., Jones & Dzhafarov, 2014). Nevertheless, the fact that decade-long debates continue about the relative merits of very different algorithmic models, such as exemplar and prototype models of category learning (Medin & Schaffer, 1978; Minda & Smith, 2002; Zaki, Nosofsky, Stanton, & Cohen, 2003), suggests that evaluation of algorithmic models could benefit from incorporating insights from other levels of analysis (e.g., Mack, Preston, & Love, 2013).

The computational level has perhaps generated the strongest enthusiasm for single-level theorizing (Anderson, 1991; Chater et al., 2006; Tenenbaum et al., 2011). The argument for forming theories at the computational level is that by focusing solely on the environment and optimality one can derive non-arbitrary theories (Anderson, 1991). However, recent critiques cast serious doubt on this assertion (Bowers & Davis, 2012; Jones & Love, 2011). Jones and Love (2011) note that by focusing solely on the environment that numerous theoretical constraints are discarded, such as those provided by physiology, neuroimaging, reaction time, heuristics and biases, and much of cognitive development. The end result is that, despite a veneer of formal elegance, computational-level theories are under-constrained and somewhat arbitrary. Unfortunately, rationality is often taken as an assumption instead of something to be tested. When a rational account fails, the theorist is free to revisit and modify ancillary assumptions about the environment and task goal until a rational account is formulated that matches behavior. Although the focus on optimality would seem to invoke favorable connections to tuning by natural selection, rational models' notions of optimality are typically impoverished because mechanism is neglected in favor of a focus on behavior. Natural selection is concerned with mechanistic considerations, such as history (natural selection works with the current gene pool, it does not "design" from a clean slate and the intermediary solutions must reproduce), and optimization involves factors typically neglected in rational analyses, such as the mechanism's energy and time requirements.

2. Top–down: Bridging from the computational level

Given the preceding criticisms of single-level theorizing and cognitive scientists' general inclination toward integration, bridging levels of analyses is likely to broadly appeal, but the question is from where to build the bridge. Two possibilities will be considered in this contribution. In this section, integration from the computational level will be considered. The next section considers integration from a mechanistic perspective. Whereas

the first approach considers the computational-algorithmic bridge, the latter focuses on an algorithmic-implementation bridge that can also be extended upward to the computational level.

Perhaps the best examples of bridging from the computational to algorithmic level involve approximate Bayesian inference (Sanborn et al., 2010; Shi et al., 2010). In approximate Bayesian inference, rather than computing the full Bayesian solution, an approximation is considered that is often easier to compute and can be viewed from a psychological perspective as embodying a capacity constraint. The basic idea is that the human cognitive architecture is not infinite capacity so it stands to reason that people use algorithms that are as rational as possible, given people's limitations. For example, Sanborn et al. (2010) consider an approximation to a rational clustering solution for categorization that better describes human performance than the full rational solution. This approximate model considers a sampled subset of solutions rather than the full posterior of solutions. Other work finds that some human biases are consistent with limitations inherent in approximation techniques (Lieder, Griffiths, & Goodman, 2013). The approximate Bayesian bridge is conceptually aligned with other work that seeks to explain human performance limitations in terms of firm capacity limits (Giguère & Love, 2013; Miller, 1956).

This approximate Bayesian bridge is appealing, but it relies on some strong assumptions that do not hold up under scrutiny. First, the starting point for this bridge is a computational-level theory. As reviewed above, computational-level theories are underconstrained. Therefore, an approximation of a computational-level theory will inherit these shortcomings. Second, people are suboptimal for numerous reasons that have nothing to do with capacity limitations. The remainder of this section will focus on this latter point, namely that approximate Bayesian accounts are likely to miss key findings at the algorithmic level that are not shaped by capacity limitations.

One challenge to the approximate Bayesian account is from work that suggests that more capacity is not always better. Less-is-more accounts have proved successful in topics in learning (Elman, 1994; Filoteo, Lauritzen, & Maddox, 2010; Maddox, Love, Glass, & Filoteo, 2008). Likewise, in many decision tasks, cognitive sophistication is not desirable and can be harmful (Beilock, Bertenthal, McCoy, & Carr, 2004; Dijksterhuis, 2004; Johnson & Raab, 2003; West, Meserve, & Stanovich, 2012; Wilson & Schooler, 1991). Contrary to the approximate Bayesian account, individual differences in people's performance in such tasks are not a function of capacity, but rather of cognitive style or approach (Frederick, 2005; Stanovich, West, & Toplak, 2013).

One fundamental limitation with the approximate Bayesian bridge is that capacity offers a single lever for capturing differences across individuals, populations, and tasks, whereas the basis for human performance is likely multidimensional. For instance, differences across development and populations suffering from neurological disease, aging, and brain insult are unlikely to reduce to a single continuum of how well the full rational solution is approximated (e.g., how many particles or samples are included in a Markov Chain approximation of the full rational solution). Likewise, introducing a dual-task manipulation will not likely make one population equivalent to another. To the extent that

a capacity account can be championed, it is likely to involve multiple capacities closely tied to notions of mechanism rather than follow from a computational-level account. Perhaps what is more likely is that different individuals and populations may differ in capacities, control structures, and key parameters all situated within a mechanism. Even lower-level work in perceptual neuroscience is coming round to the idea that noise and capacity are not the key limits to human performance (Beck, Ma, Pitkow, Latham, & Pouget, 2012).

3. Inside-out: Bridging from the algorithmic level

Unlike a computational-level theory, mechanisms make commitments to mental representations and processes. These commitments can be evaluated at the algorithmic level by comparing model predictions to behavioral measures, such as human choice and response time data. Importantly, evaluation can also be done across levels. Looking upwards, the algorithm can be compared to various computational-level theories that share similar input–output relationships with the algorithm. These comparisons are useful in identifying which aspects of the mechanism are consistent with a particular computational-level account of rationality. One useful discovery heuristic may be examining discrepancies across levels and developing explanations for why they occur. For instance, as reviewed above, one cause of discrepancy may be capacity limits. In other cases, there could be basic differences in how information is processed and updated (e.g., Sakamoto, Jones, & Love, 2008).

Looking downward, the mental representations and processes embodied in the mechanistic account can be related to implementation-level data, such as brain imaging results. This integration can go in both directions. Algorithmic accounts can be useful in understanding how the brain supports cognition, even guiding data analysis of brain measures (Anderson, Fincham, Schneider, & Yang, 2012; Davis, Love, & Preston, 2012a, b; O'Doherty, Hampton, & Kim, 2007). Additionally, brain measures can be used to select among competing algorithmic accounts if one assumes that continuity across levels of analysis is preferable to discontinuity (Mack et al., 2013). The remainder of this section will focus on the mutual constraints present at the algorithmic and implementation levels with the Conclusion section revisiting the notion of integrating mechanism with rational computational-level accounts. Although most of the examples will be drawn from the author's own work in categorization and neuroimaging, there is ample work in other domains that follows a similar progression (e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Hampton, Bossaerts, & O'Doherty, 2008).

One way mechanistic models can bear on implementation-level issues is by simulation of various populations. Mechanistic models consist of interacting components. If a bridge theory is developed relating these components to brain regions, then it is possible to make predictions about how populations with reduced function in a component will behave. Love and Gureckis (2007) related a clustering model of category learning to a learning circuit involving the medial temporal lobes (MTL) and prefrontal cortex (PFC).

Triangulating among patient, animal lesion, neuroimaging, aging, and developmental data, they were able to develop a bridge theory between model and brain. The account captured performance on certain tasks by reducing the model's ability to form new clusters in response to surprising events for certain populations. This model has subsequently correctly predicted aging results (Davis, Love, & Maddox, 2012). This work is an example of how a mechanistic model can leverage multiple constraints and data sources, which allow mechanistic models to make predictions about how varying stimuli or task conditions should affect human performance.

Although the Love and Gureckis (2007) proposal can be viewed as a capacity explanation of group differences, the theory itself details the limits of its application and does not support the position that all differences across populations reduce to a single capacity limit. Indeed, developmental simulations identify two possible capacity constraints that highlight the need for new experiments to tease apart competing accounts (Gureckis & Love, 2004).

This same theory and mechanism have been further tested in imaging studies. Neuroimaging studies offer a way to evaluate mechanistic accounts of how cognitive processing unfolds in the brain. In turn, mechanistic models offer richer ways to analyze imaging data that eclipse what is possible with standard analyses. In model-based fMRI analyses, a cognitive model is assumed to capture the psychological processes taking place in the brain. This assumption allows the model to be used as a lens on the brain imaging data. The procedure is to first fit the model to the behavioral data, then define some internal model measure of interest (e.g., recognition strength, error correction, etc.), and then observe how brain activity in different areas correlates with each model measure. Neural activity correlating with some operation in the model is suggestive that the correlated brain region may fulfill a similar function. Model-based analysis provides a theoretically satisfying method to localize mental activity in terms of the operations of a mechanism. Without a model, one is left with standard analyses that compare activations elicited by different experiment conditions (e.g., viewing a face vs. a house).

What follows is a brief overview of work in categorization that adopts an integrative model-centric approach (please see the original contributions for full details). Using this model-based approach, Davis, Love, et al. (2012a) fit the SUSTAIN clustering model (Love, Medin, & Gureckis, 2004) to behavioral data from a category learning experiment in which there was an imperfect rule for categorizing novel stimuli into one of two categories. Subjects had to learn to apply the rule to rule-following items and refrain from applying the rule to exception items. The basic findings from rule-plus-exception studies is that people make more errors on exception items during learning, but subsequently show better recognition memory for these items (cf. Nosofsky, Palmeri, & McKinley, 1994; Sakamoto & Love, 2004). The model provided a way to "observe" non-observable mental activities during learning, such as recognition and error correction. On each trial, a measure of the model's recognition strength at stimulus onset (at the beginning of the trial) and error correction at the end of the trial when feedback is provided was calculated. Recognition strength was modeled as the sum of cluster activations, whereas error correction was related to weight change following corrective feedback.

These two model measures are shown in Fig. 1. As predicted, these model operations capture the activity pattern of the MTL. As learning progresses across trials, the MTL ramps-up activity (particularly for exception items) at stimulus onset and ramps-down activity at feedback. The model makes these psychological processes related to recognition and error correction observable in the brain, confirming Love and Gureckis's (2007) predictions about the role of the MTL in category learning. Interestingly, a standard analysis contrasting regions that show greater activation for exception than rule-following items did not reveal any significant differences. This negative result indicates that the MTL is involved in the dynamic learning and recognition processes characterized by the clustering model, rather than statically responding to different item types. These results provide insight into the computational role of the MTL and also support the clustering account of category learning for this task.

This work demonstrates that two processes occurring in different phases of the same trial can be successfully localized by using a mechanistic model to guide the analysis. Interesting, two simultaneously occurring processes can also be localized by defining two different model measures for the same phase of a trial and considering which brain

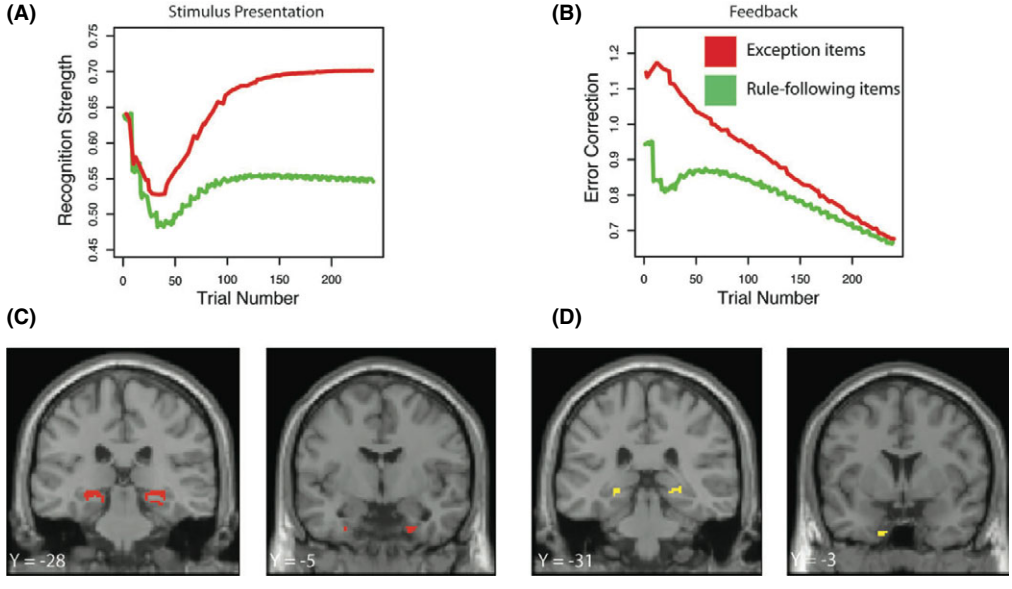

Fig. 1. Illustrations of recognition strength (A) and error correction (B) measures derived from the SUSTAIN clustering model that were used as predictors of brain activation during stimulus presentation and feedback. Recognition strength is the sum of the total cluster activation for a given stimulus/trial, and error is the absolute value of the model's output error on a given trial. Below the measures are the corresponding statistical maps associated with each regressor. Activation during stimulus presentation is presented in red and activation during the feedback period in yellow. (C) MTL regions exhibiting a significant ($p < .05$, FDR corrected) correlation between activity during the categorization period and the predicted recognition strength measure. (D) MTL regions exhibiting a significant ($p < .05$, FDR corrected) correlation between during feedback and the predicted error correction measure.

regions best track each. Using this approach, Davis, Love, et al. (2012b) were able to confirm predictions about differences in regions that support recognition and those related to cluster entropy (i.e., uncertainty about cluster membership). Such hypotheses would be impossible to test in fMRI without a model to tease apart simultaneously occurring processes within a single imaging data stream.

Thus far, I have detailed how mechanistic models can be used to help identify mental process occurring through time in learning tasks. Multivariate model-inspired analyses can also be used to examine representation of learned categories in the brain (Davis, Xue, Love, Preston, & Poldrack, 2014). Davis et al. confirmed that the representations developed in the MTL are analogous to those acquired by the SUSTAIN clustering model during rule-plus-exception learning. Davis et al. performed a similarity analysis of the MTL's representational space toward the end of the category learning task. The brain response (across voxels in the MTL) was recorded for the different learning items, and the similarity of these brain representations was assessed by correlation. This neural similarity analysis was subjected to a multidimensional scaling procedure (see Fig. 2). As a result of learning, the brain (like the mechanistic model) remaps the items into a new space that segregates items by category and places the exception items in the center of the representational space, which explains why these items both generate more errors (because they are more confusable with opposing category items) in learning and are

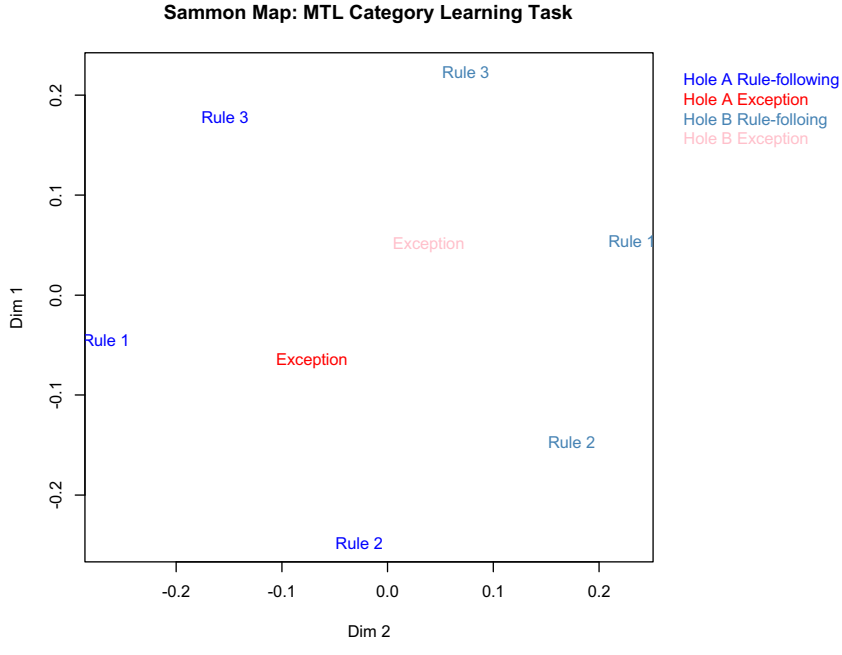

Fig. 2. A Sammon Map (a type of multidimensional scaling) of between-stimulus pattern similarities at the end of learning during the category learning task. Rule-following items are denoted by "Rule," whereas exception items are denoted by "Exception."

recognized more reliably (because they are more similar to other item representations and thus more familiar). Early in learning, the MTL makes no category or item distinctions.

The aforementioned work demonstrates that mechanistic models are very useful for understanding how the brain supports mental processes and representations. However, is the relationship between algorithm and implementation a two-way street (cf. Turner et al., 2013)? In other words, can we use brain-level data to determine the best psychological model? Recently, Mack et al. (2013) developed a method that allows one to evaluate which set of competing cognitive models is most consistent with brain response. Whereas the aforementioned work assumed the correct model, this method uses brain response to adjudicate among competing models, which is particularly useful when behavioral data alone are insufficient for discriminating between competing accounts. The method works by using machine learning techniques to evaluate the mutual information between brain and model state changes. Models are favored to the extent that their internal state changes are predictable by brain response.

Using this technique, Mack et al. found that brain response in the classic Medin and Schaffer (1978) category learning task was more consistent with exemplar (i.e., concrete) than prototype (i.e., abstract) representations of categories. Behavioral data alone could not discriminate between the models as both models did an excellent job fitting the choice data, which is not surprisingly given that the field has debated for decades about whether the exemplar or prototype model is the best account for this task (Medin & Schaffer, 1978; Minda & Smith, 2002; Zaki et al., 2003). This novel technique also allows for finer grain questions to be answered, such as which regions support which model components. This method and accompanying results demonstrate the advantages of working with mechanistic models. These models can leverage a variety of data sources to constrain theory development, including using brain response to tease apart competing algorithmic-level accounts.

4. Conclusion

In this contribution, I considered limitations in theorizing at any one of Marr's levels of analysis. Then, two avenues for integration across levels were discussed. The first approach was "top-down" in that it attempted to bridge from the rational computational level to the algorithmic level. In this case, the strategy is to assume that all suboptimalities in behavior arise from capacity limitations that can be modeled as approximate Bayesian inference. In other words, people are computing an approximation to the "correct" rational solution. One major problem with this path to integration is that the computational-level base for the bridge is under-constrained (Bowers & Davis, 2012; Jones & Love, 2011) and therefore the arbitrariness of the rational account is inherited by the approximate Bayesian account. A second major problem is that people are suboptimal for many reasons other than capacity limitations. Thus, the approximate Bayesian approach will not capture many findings at the algorithmic level.

A second path to integration is to use algorithmic-level models to better understand implementation-level findings and vice versa. This strategy is extremely promising as it exploits mutual constraints across levels. In the examples considered above, algorithmic models were used to interpret brain imaging findings these were used to test between competing algorithmic models. Representations and processes of mechanistic models were evaluated at both levels of analysis.

In principle, nothing prevents the inside-out approach from being extended upward to the computational level. Indeed, the Davis, Love, et al. (2012b) study used Sanborn et al.'s particle filter model (an approximate Bayesian model) to simultaneously measure processes related to recognition and entropy across cluster membership. To be clear, Davis et al. used the Sanborn et al. model in a mechanistic fashion (i.e., model operations were treated and evaluated as psychological processes and representations) as opposed to as a rational model justified by formal analysis of the learning problem. In the Davis et al. work, because of the mathematical formulation of the mechanistic model, it is straightforward to relate the modeling at the algorithmic level to a computational-level account, thus spanning all three levels of analysis.

This integration across levels raises questions about the necessity of Marr's levels. Essentially, the findings and methods I describe blur Marr's distinctions. Indeed, the method described in Mack et al. removes the separation between the algorithmic and implementation levels. The only assumption necessary to blend these levels is that one should prefer algorithmic models that supervene in some discernible way on the implementation level. This assumption amounts to little more than saying that mental activity depends on a physical substrate, which should not be controversial. One could conclude that Marr's levels have outlived their usefulness and are self-imposed barriers to investigation. Nevertheless, these levels may have value in making researchers aware of their intended contribution and the data sources relevant to theory evaluation.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Anderson, J. R., Fincham, J. M., Schneider, D. W., & Yang, J. (2012). Using brain imaging to track problem solving in a complex state space. *NeuroImage*, *60*(1), 633–643.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*(5), 543–564.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Beilock, S. L., Bertenthal, B. I., McCoy, A. M., & Carr, T. H. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review*, *11*(2), 373–379.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Where next? *Trends in Cognitive Sciences*, *10*(7), 292–293.

Cooper, R. P., & Peebles, D. (2015). Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, *7*(2), 243–258.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, England: Clarendon Press.

Davis, T., Love, B. C., & Maddox, W. T. (2012). Age-related declines in the fidelity of newly acquired category representations. *Learn & Memory*, *19*(8), 325–329.

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based FMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273.

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 821–839.

Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience*, *34*(22), 7472–7484.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology*, *87*(5), 586–598.

Elman, J. L. (1994). Implicit learning in neural networks: The importance of starting small. In *Attention and performance XV: Conscious and nonconscious information processing* (pp. 861–888).

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, *21*(3), 415–423.

Franz, S. I. (1912). New phrenology. *Science*, *35*, 321–328.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Giguère, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences USA*, *110*(19), 7613–7618.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Gureckis, T. M., & Love, B. C. (2004). Common mechanisms in infant and adult category learning. *Infancy*, *5*, 173–198.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences USA*, *105*(18), 6741–6746.

Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision*, *91* (2), 215–229.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability of major modeling schemes for choice reaction time. *Psychological Review*, *121*(1), 1–32.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences*, *34* (4), 169–188; disuccsion 188–231.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2013). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, *26*.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective & Behavioral Neuroscience*, *7*(2), 90–108.

Love, B. C., Medin, D. L., & Gureckis, T. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, *111*, 309–332.

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*, 2023–2027.

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition*, *108*(2), 578–589.

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

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 275–292.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & Cognition*, *36*(6), 1057–1065.

Sakamoto, Y., & Love, B. C. (2004). Schematic Influences on Category Learning and Recognition Memory. *Journal of Experimental Psychology: General*, *33*, 534–553.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

Utall, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. London: MIT Press.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, *103*(3), 506–519.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1160–1173.