How could a rational analysis model explain?

Samuli Reijula (samuli.reijula@helsinki.fi)
TINT / Social and Moral Philosophy
PO BOX 24, 00014 University of Helsinki, Finland

Abstract

Rational analysis is an influential but contested account of how probabilistic modeling can be used to construct non-mechanistic but self-standing explanatory models of the mind. In this paper, I disentangle and assess several possible explanatory contributions which could be attributed to rational analysis. Although existing models suffer from evidential problems that question their explanatory power, I argue that rational analysis modeling can complement mechanistic theorizing by providing models of environmental affordances.

Keywords: probabilistic modeling; rational analysis; scientific explanation; mechanism; affordance

1. Introduction

During the past two decades, probabilistic modeling has become one of the most visible strands of cognitive modeling alongside connectionism, dynamical systems, and rule-based approaches. Curiously, against the general trend in the psychological sciences where theorizing is increasingly anchored in neuroscience findings, probabilistic modeling of higher cognition has been a characteristically top-down endeavor. Without making any substantial commitments about the underlying cognitive mechanisms, probabilistic modeling has been applied to complex aspects of human cognition, which still largely remain beyond the reach of mechanistic research methods. Models of human memory, categorization, causal learning, concept learning, and conditional inference, to mention a few applications, often show an impressive fit to empirical data, and the novel analyses of cognitive capacities provided by the models appear to have shed new light on the nature of the studied phenomena.

However, how does that shedding light actually occur – how do such computational probabilistic models explain? Although probabilistic modeling, in principle, does not rely on any particular method of explanation (and not all models aim to be explanatory), modelers often refer to the idea of rational analysis as the account of how and why their models help us understand the mind (Anderson 1990; Oaksford & Chater 2007). The striking claim made by rational analysis (RA) modelers is that by treating higher cognitive capacities as forms of inductive inference, we can predict behavior, and explain a lot about human cognition without making any assumptions about the underlying representations and processes. This agnosticism about implementation is typically justified by making reference to a rationality assumption: We know that human agents tend to be well-adapted to their environment, and hence a careful analysis of the cognitive task encountered by the mind, coupled with an assumption of the optimality of human behavior in the task, results in a putatively powerful methodology of prediction and explanation.

However, it is a widely-held view in the philosophy of science that explanations, also in the cognitive sciences, should track causal mechanisms, and the way that RA purports to sidestep the evidential and explanatory problems arising from the causal complexity of cognition has given rise to a strongly polarized debate (see, e.g., Jones & Love 2011). On the one hand, the way that the new mathematical methods in probabilistic modeling can capture the interplay of structure and learning in human thought has led to the emergence of an exciting research paradigm. On the other hand, the proponents of non-mechanistic (or even non-causal) explanation need to show when and how it is that such models genuinely explain rather than only redescribe or merely formally unify various phenomena (see Colombo & Hartmann 2017). Failing to do that, rational analysis could simply be seen as the last breath of the autonomist dream of studying the mind independently from the brain.

The goal of this paper is to advance the debate by disentangling various explanatory contributions which can be attributed to RA models. By relying on the influential contrastive-counterfactual account of explanation, I distinguish between three possible explanatory contributions: Uncovering (a) constitutive dependencies in cognitive systems (i.e. dependencies between parts and wholes), (b) environment-behavior dependencies, and (c) environment–optimal behavior dependencies.[1] I argue that often the option (c) best describes the nature of the new understanding provided by RA models: In many cases, RA models should be interpreted as being explanatory not of human behavior as such, but of environmental affordances. Consequently, well conducted modeling of environmental affordances can complement mechanistic theorizing by providing resources for understanding the possible space of behavior of agents.

2. Probabilistic modeling and rational analysis

2.1 Procedure of rational analysis

The idea of rational analysis modeling dates back to John Anderson's work on human memory and categorization in *The Adaptive Character of Thought* (1990). Having already worked on his ACT* cognitive architecture, the new methodology put forward in the book reflected Anderson's

[1] To be clear, probabilistic models are also used for purposes other than explanation (e.g., prediction, hypothesis generation). This paper, however, only examines their explanatory import.

increasing worries that the research methods of the time could not really uncover cognitive mechanisms. Lacking a clear picture of what it is that cognitive mechanisms do (i.e. what the psychological explananda are), the available evidence of cognitive processes and their neural implementation was insufficient to uncover the mechanistic architecture of the human mind (Anderson 1990, pp.23–26). Compared to bottom-up research strategies, rational analysis begins from the other end:

> *[…] We can understand a lot about human cognition without considering in detail what is inside the human head. Rather, we can look in detail at what is outside the human head and try to determine what would be optimal behavior given the structure of the environment and the goals of the human.* (Anderson 1990, p.3)

According to Anderson, careful mathematical modeling of the environment and task structure combined with an assumption about the optimality of human behavior leads to a new self-standing research strategy for understanding the mind: *"As this book is evidence, a rational analysis can stand on its own without any architectural theory"* (ibid.). By providing a precise model of what the mind does, rational analysis can constrain the search space for cognitive mechanisms, and, putatively, put the scientific study of the mind on a firm foundation.

This view of the role of computational modeling immediately brings to mind Marr's (1982) account of multi-level theorizing. However, whereas Marr provides no systematic account of how computational-level theories are to be constructed, RA modeling has predominantly proceeded according to the six-step cycle proposed by Anderson (1990, p.29):

1. Specify precisely the goals of the cognitive system
2. Develop a formal model of the environment to which the system is adapted
3. Make minimal assumptions about computational limitations
4. Derive the optimal behavior function, given items 1 through 3
5. Examine the empirical evidence to see whether the predictions of the behavior function are confirmed
6. Repeat, iteratively refining the theory

These steps embody an account of how a large part of probabilistic cognitive modeling is done. However, two further assumptions of RA should be made explicit. First, the derivation of optimal behavior in steps 2-4 typically employs *probability calculus* (not logic) as the normative baseline theory of rational behavior. Secondly, the link between model predictions (step 4) and observed behavior of humans (step 5) is formed by an assumption about the *optimality of the observed behavior* (see quoted passage above).

Below I illustrate this process with an example. However, a comment on the status of the approach in cognitive science is in place: Obviously, not all probabilistic modelers endorse the rational analysis framework (see Brighton & Gigerenzer 2008; Danks 2015; Frank 2013). Focusing on RA is useful for two reasons, however. The approach is undeniably influential, and its core commitments have been endorsed by a large group of well-known modelers (e.g., Anderson 1990; Oaksford & Chater 1994; Griffiths & Tenenbaum 2009). A further advantage of focusing on RA has to do with the fact that often the theoretical commitments of mathematical modelers can be hard to pin down. In some cases, the ambiguities are surely due to the modelers themselves not being clear about where their commitments (about how to understand explanatoriness, optimality, etc.) lie. Rational analysis is a clear account of the conceptual foundations of probabilistic cognitive modeling, and provides a starting point, or at least a foil, for explicating such commitments.

To illustrate the rational analysis process, I now briefly introduce Mike Oaksford and Nick Chater's (1994, 2007) analysis of the Wason selection task.

2.2 The information gain model

Wason selection task is one of the most famous laboratory experiments discussed in the literature on human rationality. In the original form of the task, participants are given four cards, each of which has a letter on one side and a number on the other. The participants' task is to determine whether the rule *"If there is a vowel on one side of the card* (p)*, then there is an even number on the other side* (q)*"* holds. More precisely, they are asked to select those cards which must be turned over to discover whether the rule is true or false. The famous finding from the task and its several replications is that only a small minority (less than 10%) select the correct cards (vowel, odd number) corresponding to the falsifying instance. That is, judged in the light of logic, most participants fail to perform in a rational way.

Oaksford and Chater (O&C) challenge the irrationality claim by arguing that logic-based theories of inference and rationality misrepresent the participants' behavior in the task. O&C's own *information gain model* suggests that people's apparently irrational way to test a hypothesis should actually be seen as optimal strategy for uncertainty reduction. The gist of O&C's reinterpretation is that instead of engaging in deductive reasoning, participants interpret the task as an inductive one. They do not try to falsify the rule, but instead they try to determine which of two hypotheses holds:

a) Independence model, $\mathbf{M_I}$: $P(q|p) = P(q)$ or
b) Dependence model, $\mathbf{M_D}$: $P(q|p)$ is high, higher than $P(q)$.[2]

Being initially equally uncertain about both hypotheses, participants aim to reduce this uncertainty as much as possible by turning as few cards as possible.

[2] In Oaksford & Chater 2007, $P(q|p)$ was set to 0.9. See ibid. for the underlying account of conditional inference and for the mathematical details.

The rational analysis proposed by O&C relies on three core principles:

1) Higher cognition can be modeled as probabilistic (Bayesian) computation.
2) The likelihoods and prior probabilities required by the model can be acquired from the analysis of the environment structure.
3) Behavior of human agents constitutes an optimal response to the task.

The model is constructed roughly as follows. To formalize the idea of uncertainty reduction, O&C adopt the optimal data selection paradigm, and interpret uncertainty reduction as optimizing *expected information gain*. Information gain $I_g(D)$ from turning over a card (D) is defined as $I(M_i|D) - I(M_i)$ where the Shannon information[3] $I(M_i)$ can be derived from the probabilities of the hypotheses before and after observing data, $P(M_i)$ and $P(M_i|D)$. The required posteriors can be obtained by the Bayes' rule from the likelihoods $P(D|M_i)$ and the prior probabilities of the hypotheses. Reflecting initial ignorance, the priors were set to 0.5 and hence the rest of the crucial model specification is built into the likelihoods, which reflect the nature of the four-card task. Oaksford and Chater (1994) show in detail how the required likelihoods can be read off the contingency tables describing the two hypotheses.

From these derivations, it follows that the base rates of p and q have a central role in determining which behavior is optimal. They describe how frequently positive instances of the antecedent and consequent of the rule appear in the environment. Qualitatively, the expected information gain from each of the four cards turns out to depend on the base rates P(p) and P(q) in the following way:

P(q) is small	→ p card is informative
P(p) is large	→ not-q card is informative
P(p) and P(q) are small	→ q card is informative (not-p card is not informative)

How should these base rates, then, be determined? Instead of attempting to somehow measure them in relevant environments for different kinds of rules, O&C cite various intuitively plausible justifications for their *rarity assumption*: Relying on the observation that categories in language cut the world quite finely, and that properties that figure in causal relations tend to be rare, the assumption states that, generally, P(p) and P(q) are small in most situations. Under rarity, O&C conclude, the q card is more informative than the *not-q* card. Hence, the model suggests that the highest expected information gain is achieved by turning over the p and q cards, exactly as the majority of the participants do. In fact, with P(p)=0.22, P(q)=0.27, the model shows a very good fit to experimental data from the Wason task. Hence, by changing the normative model of rational behavior, O&C were able to explain away irrationality, and to show that participants' behavior in experiments is actually close to optimal.

The model has received critical attention in the literature (see Oaksford & Chater 2009), but it serves our current purposes well. The model specification and the modeling assumptions are conceptually on a par with those in more complex Bayesian models: The complexity often pertains to the number of variables involved, the structure and generation of the hypothesis space, and in many cases advanced numerical methods are needed for solving the model. These mathematical sources of complexity do not, however, change the fundamental conceptual architecture of a model. What is common to all RA models is that none of their main components (hypothesis space, likelihood function, priors) are interpreted in a psychologically realistic way as mental representations (Jones & Love 2011). Instead, they stand directly for properties of the environment. Furthermore, empirical data about the properties of human cognition is not fed into the model specification to calibrate or to constrain the model. Instead, behavioral data enters only in step 5 of RA (see above) as a means for testing model predictions. In this sense, the information gain model is an illuminating example of the theoretical and conceptual assumptions made in rational analysis modeling.

3. What rational analysis models fail to explain

There is no consensus in philosophy (or in the sciences) about what scientific explanation is, or what makes a theory explanatory. However, a shared starting point for many accounts of scientific explanation has been to distinguish explanation from other epistemic activities (e.g., description and prediction) by pointing out that explanations offer information of a specific kind. Explanations show *how* or *why* something happened or obtains. According to an influential approach (Woodward 2013), the knowledge that allows one to answer such questions concerns change-relating counterfactual dependencies between the relata in the explanation, the *explanans* and the *explanandum*. That is, explanations show how (the state of) some things depend on (the state of) other things.

This contrastive-counterfactual account of explanation suggests that explanatory information has generally the following form:

{CC} *y[y'] because of x [x']* (variable Y takes the value y instead of y' because X has the value x instead of x')

According to the contrastive-counterfactual account, being able to explain means that one is able to correctly answer *what-if-things-were-different questions*, i.e. questions about how changes in explanantia variables influence the state of the explanandum variable. In addition to being a sufficiently general account of explanation, the contrastive-counterfactual account suits the purposes of this article well, because it does not necessarily tie the notion of explanation

[3] Uncertainty $I(M_i)$ given n mutually exclusive and exhaustive hypotheses, is $-\sum_{i=1}^{n} P(M_i) \log_2 P(M_i)$.

to that of causation. That is, although the 'because' in {CC} is typically understood as referring to a causal dependency, the account does not rule out the possibility of there being also non-causal explanations (Woodward 2013; Pincock 2015; Rice 2015): If a suitable analysis of invariant dependency in non-causal contexts (e.g., for mathematical dependencies) can be found, the contrastive-counterfactual account can be applied to non-causal explanations as well. Hence, the account of explanation casts the net wide enough to give RA models a fair chance of being explanatory.

A further advantage of treating explanations as answers to questions is that it allows us to sharpen the *explananda*, i.e. to make more precise the possible explanatory claims arising from RA models. I suggest that there are at least three different kinds of objective dependencies that RA models could be said to track: (1) constitutive dependencies between parts and wholes, (2) environment-behavior dependencies, and (3) environment–optimal-behavior dependencies. In the rest of this section, I argue that often RA models do not have genuine explanatory import with respect to the two first kinds of dependencies. The more promising third option is discussed in section 4.

3.1 Constitutive what-ifs

The notion of mechanism has acquired a central position in the philosophical debates on scientific explanation. A clear expression of the mechanistic viewpoint has recently been given in the *model-to-mechanism mapping (3M) requirement* by Kaplan and Craver (2011). According to the requirement, dynamical and mathematical models in systems- and cognitive neuroscience can be explanatory only if there is a mapping between elements in the model and elements in the mechanism for the phenomenon. As the example discussed above suggests, rational analysis models provide no such mapping. They are agnostic about algorithmic and implementation level details, and intentionally so. They clearly do not track constitutive dependencies. Does this mean they cannot be explanatory?

As Kaplan and Craver themselves admit, their argument ultimately relies on shared norms about explanatoriness in the neuroscience community, and their account of explanation as uncovering multi-level mechanisms reflects these norms. If such norms do not hold among probabilistic cognitive modelers, it is not obvious why they, based on this argument alone, should abide by the 3M requirement.

The contrastive-counterfactual account suggests a more positive reply to Kaplan and Craver's argument: RA models obviously do not provide information about constitutive and causal dependencies in multi-level mechanisms, but this does not rule out the possibility of them tracking some other kinds of objective dependencies, for example, those holding between relata described in purely computational-level terms. Furthermore, a proponent of RA need not (and should not) claim that adding mechanistic detail never improves a computational explanation. To defend the explanatoriness of RA models, a far weaker claim suffices, one stating that it is possible to learn about objective explanatory dependencies without always relying on information about cognitive mechanisms.

3.2 Environment–behavior what-ifs

A second kind of explanatory question answered by an RA model could be: *"How would the behavior of the cognizer change when the cognitive task changes in some particular way?"* That is, the model could uncover objective dependencies between properties of the environment and the behavior of cognizers. For example, O&C's model can be used to derive predictions about what the behavior of the participants in the Wason task would be, were P(p) and P(q) to take some range of values.

It is here that the optimality assumption of RA becomes crucial. To predict how human behavior would change in response to changes in the task, without knowing anything about the algorithms and processes producing the behavior, RA relies on the assumption that humans are well-adapted to their environments: If we assume that human behavior is close to optimal across a large variety of environments, the predictions derived from the RA model (step 4 of the RA procedure) should in fact apply to that behavior. Optimality forms the link between the normative theory and observed behavior.

Given that human (ir)rationality has been the topic of a longstanding debate in philosophy and psychology, it is not surprising that the optimality assumption has drawn a lot of criticism (Jones & Love 2011). Although proponents of RA are correct in arguing that some degree of rationality of target behavior is required for us to even perceive it as intentional action, such modest levels of rationality hardly license the strong optimality assumptions of RA models. Neither do evolutionary arguments provide support for strong optimality claims: Natural selection is a source of design and adaptedness, but not necessarily of globally optimal solutions – merely a local comparative advantage is sufficient for evolutionary solutions to survive.

Being aware of these problems, proponents of RA have avoided evolutionary defenses of the optimality assumption. Instead, they often justify optimality by relying on analogies to behavioral ecology and economics, where similar assumptions are commonly made (Chater et al. 2003). However, such analogies break down due to a crucial dissimilarity between these fields. Unlike in cognitive science, both in biology and economics the rationality claims typically concern aggregate behavior, not that of individual agents. Hence, I do not see how appealing to economics or biology could be a viable way to justify optimality assumptions in RA modeling.

These problems with the general defenses of the optimality assumption suggest that perhaps optimality should be examined more locally. Now, what kind of evidence should be obtained to justify the optimality claim in the case of a particular cognitive task? It seems that to support a claim about there being an objective dependency between environment and behavior, we should gather data about human behavior in a task *across a range of parameter values*

describing various different environmental states. In other words, if human behavior fits the predictions of the model across a range of conditions, that would appear to be rather strong evidence of optimal performance.

However, the existing RA models rarely make use of such cross-environmental data. First of all, many models do not rely on any actual measurements of environment parameters. Instead, they use plausible-sounding assumptions or analogies. For example, Anderson (1990, ch. 2) relied on data about library borrowings to model usage of memory structures, and Griffiths et al. (2007) use Google PageRank to predict fluency of recall. Models devoid of good quality empirical data should be considered as toy models (at best), incapable of uncovering the actual properties of cognitive environments. Furthermore, Marcus and Davis (2013, table 1) suggest that Bayesian modelers have been selective in choosing the results that they report from experimental tasks, only reporting results where human behavior follows the model predictions and ignoring cases where behavior is not optimal.[4]

That said, in the large literature on the information gain model, predictions from the model have been tested against human performance under different base rates and different framings of the task (e.g., descriptive vs. deontic; Oaksford & Chater 2007, Ch.6). Although the empirical findings remain inconclusive, such systematic variation of the task parameters should be used to produce evidence of a robust explanatory environment-behavior dependency.[5]

4. Rational analysis and the logic of the situation

Finally, let us examine the epistemic value of an RA model if we drop the optimality assumption. Assume that we have a model with a (i) well-specified task structure, (ii) parameter values based on measurements of the environment, and (iii) an empirically informed account of computational costs and cognitive limitations. What such a model could do is it could link combinations of parameter values to best possible behavioral choices in those situations. Is this not a kind of objective change-relating dependency? However, consider what the relata of such a dependency are. The model tells what the optimal behavior would be, given a particular combination of environmental conditions and computational limitations. Such counterfactuals do not say anything about actual human behavior. Instead, they can be seen as increasing our understanding of the environmental affordance, or, the logic of the situation (Popper 1963).[6]

What mathematical models of affordances (of the opportunities that the environment offers for the agent) can help us understand is the possible space of action for cognitive agents. Models of affordances show what a hypothetical rational agent would do in different situations. For what kind of purposes could such information be useful? First, were we to design artificial cognitive systems with a particular cognitive task in mind, these systems should approximate the optimal behavior specified by the model. For example, in the selection task, *if* we are interested in reducing our uncertainty, O&C's model tells us something non-trivial, i.e. which information sources to examine given the base rates of p and q.

Secondly, as in economics, rational models can obviously act as normative baselines to which human behavior can be compared. As Sloman & Fehrbach (2008) argue, often it is just as interesting to discover that behavior does not conform to the rational norm as to see that it does. Finding out when and how complex systems malfunction is often an efficient way to learn about the underlying processes.

However, in neither one of these cases are RA models used to directly explain human behavior. Instead, the model functions as an inferential aid which helps to chart the possible space of action for agents, when faced with a particular task. Herein lies perhaps the hardest evidential problem faced by rational analysis. How do we know what the mind really does in some situation; where do the functional hypotheses in step 1 of RA come from? For example, how would O&C defend their Bayesian construal of the selection task against an adamant falsificationist? The currently available empirical evidence can hardly decide the issue: Where O&C see optimal behavior, the falsificationist sees well-known inferential blunders. Marcus and Davis (2013) have argued that similar problems of model selection plague several other RA models as well.

The difficulty seems to come down to the fact that the cognitive tasks and the affordances available to an organism depend on its "life space" – not the physically objective world in its totality, but reality filtered through the organism's needs, drives and perceptual apparatus. Therefore, we should not think that the researcher's intuitions are necessarily a reliable guide to what the tasks faced by different aspects of the human cognition really are. Ad-hocness in task specification, in turn, raises serious worries about the *relevance of RA modeling*: Constructing detailed mathematical models of potential affordances is of little interest unless such affordances can be shown to be ones actually offering themselves to the human mind.

This worry suggests that the six-step rational analysis modeling cycle introduced in section 2.1 should not proceed independently from knowledge originating from mechanistic research: As both the connectionist rivals of RA and proponents of multi-level mechanistic explanation have argued, functional hypotheses (step 1 of RA) in cognitive science must be formulated in an iterative process between bottom-up and top-down research strategies (see McClelland et al. 2010; Bechtel & Richardson 2010). In particular, knowledge of perceptual capacities and embodiment

[4] See Goodman et al. (2015) for the modelers' response.

[5] See Griffiths & Tenenbaum (2006) for an empirical study that attempts to directly test the optimality assumption.

[6] As an anonymous referee pointed out, also the dynamical models used in ecological psychology are often understood as formalizations of affordances. This calls for a systematic comparison between the two modeling paradigms.

(informing step 2), as well as of the computational constraints of organisms (step 3) mostly originate from the bottom-up research on the mind-brain, and this knowledge should be allowed to constrain RA models. In this sense, Anderson's and O&C's claims about the self-standing explanatory role of RA are not vindicated by my analysis.

However, neither can bottom-up research strategies stand on their own, or at least they fail to reach high enough. The discussions on mechanistic explanation often have a reductionist bias, and understanding the environments within which cognitive mechanisms function has not received sufficient attention. Here RA models can complement mechanistic theories of cognition by providing precise mathematical models of the task and the environment. For example, as Chater et al. (2003) point out, a correctly formulated rational analysis can show why it is that some simple heuristic can be successful in solving a computationally complex task.

5. Conclusions

I have argued that given a sufficiently broad account of scientific explanation, there are several possible ways in which probabilistic modeling could increase our understanding of the mind. However, the strictly computational-level approach embodied in the six-step formula of rational analysis has led to theorizing which often fails to reliably uncover genuine explanatory dependencies. The shortcomings of RA are evidential in nature: The data, and the way it is used in model construction, often cannot support the counterfactual inferences needed explaining human behavior.

My new proposal about the epistemic role of RA models is that they can be understood as models of environmental affordances. Interpreted in this way, the models do not actually provide information about the mind works, or even hypotheses about actual cognitive functions (cf. Marr 1982; Zednik & Jäkel 2014). Instead, they help to chart the possible cognitive space of action for an organism. The nature of the explanatory contribution of such information is best worked out as a part of a non-reductionist mechanistic research programme.

References

Anderson, J. (1990). *The Adaptive Character of Thought*. Hillsdale, JN: Lawrence Erlbaum Associates.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity.* The MIT Press.

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: harmony or dissonance? In Chater & Oaksford (eds.) *The Probabilistic Mind*. Oxford University Press.

Chater, N., et al. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes,* 63–86.

Chater, N., & Oaksford, M. (eds.) (2007). *The Probabilistic Mind.* Oxford: Oxford University Press.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *British Journal for the Philosophy of Science,* 68, 451-484.

Danks, D. (2015). *Unifying the Mind*. MIT Press.

Frank, M. (2013). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*, 128, 417–423.

Goodman, N. Frank, M., Griffiths, T., Tenenbaum, J., Battaglia, P., & Hamrick, J. (2015). Relevant and robust: A response to Marcus and Davis. *Psychological Science*, 26, 539-541.

Griffiths, T., Steyvers, M., & Firl, A. (2007). Google and the mind. *Psychological Science*, 1069–1076.

Griffiths, T., & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 767-773.

— (2009). Theory-based causal induction. *Psychological Review*, 661–716.

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? *Behavioral and Brain Sciences*, 34, 169-231.

Kaplan, D., & Craver, C. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78, 601-627.

Marcus, G., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24, 2351–2360.

Marr, D. (1982/2010). *Vision*. W.H. Freeman/MIT Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.

— (2007). *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford University Press.

— (2009). Precis of Bayesian Rationality. *Behavioral and Brain Sciences*, 69–120.

Pincock, C. (2015). Abstract explanations in science. *British Journal for the Philosophy of Science* 66, 857-882.

Popper, K. (1963). Models, instruments, and truth. Manuscript. Karl Popper Collection at the Hoover Institution Archives at Stanford University.

Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. *Noûs* 49, 589-615.

Sloman, S., & Fehrbach, P. (2008). The value of rational analysis: as assessment of causal reasoning and learning. In *The Probabilistic Mind.*

Woodward, J. (2013). Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society Supplementary Volume,* lxxxvii: 39–65.

Zednik, C., & Jäkel, F. (2014). How does Bayesian reverse-engineering work? In P. Bello et al. (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 666-671.