2008

Rational Analyses, Instrumentalism, and Implementations

David Danks
Carnegie Mellon University, ddanks@cmu.edu

Rational Analyses, Instrumentalism, and Implementations

David Danks

Department of Philosophy, Carnegie Mellon University; and

Institute for Human & Machine Cognition

To appear in:

N. Chater & M. Oaksford, (Eds.), *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.

Address correspondence to:

David Danks
Department of Philosophy
135 Baker Hall
Carnegie Mellon University
Pittsburgh, PA 15213
ddanks@cmu.edu
Tel: (412) 268-8047
Fax: (412) 268-1440

Acknowledgements: Early versions of these ideas were presented in talks at "The Probabilistic Mind: Prospects for Rational Models of Cognition" workshop held at the Gatsby Computational Neuroscience Unit (University College, London), and at the Center for the Philosophy of Science (University of Pittsburgh). Thanks to audiences there, and to Nick Chater, Clark Glymour, Tom Griffiths, Edouard Machery, Josh Tenenbaum, and Annika Wallin for helpful discussions.

Introduction

Rational analyses provide explanatory models of cognition, and I focus here on the nature of that explanation. I argue first that many—and perhaps even most—rational analyses fail to meet the evidential standards required to provide genuine explanations of the sort commonly attributed to them. Put simply, rational analyses (or at least, the actual instances that we usually have) are not as powerful as is normally thought. Second, I argue that there should nonetheless be an expansion of rational analysis to domains that have been thought off-limits. That is, rational analyses can be much more useful than is normally thought. In sum, this chapter argues for a shift in the practice of rational analysis to more widespread application, but with more careful consideration of their actual power. Rational analyses are an important tool that should be used more widely, but with more care about what they actually yield.

A platitude about the sciences is that they are partly (or primarily) in the business of providing explanations. Predictions and data summaries are important and practically useful, but we expect our best scientific theories to tell us something about *why* some phenomenon occurred. The cognitive sciences are no different in this regard: we want an explanation for people's behavior, not just a prediction or retrodiction of it. Rational analyses purport to offer a different type of explanation from that normally found in cognitive science, since the explanation is supposed to be agnostic about underlying cognitive mechanisms. The implication, as I will argue in the second section, is that many current rational analyses are "simple" instrumentalist theories: they offer data summaries and potentially predictions, but essentially no additional explanatory power.

Instrumentalist theories are not necessarily to be avoided in all cases. I argue in the third section that there are conditions in which instrumentalist theories—and rational analyses in

particular—can be valuable, but that we should not be content with such theories. A constant goal (perhaps suspended for a particular project, but still remaining in the background) should be to determine the cognitive mechanisms that underlie particular rational analyses. Understanding this goal requires an understanding of the nature of 'implementation,' and so I turn (in the fourth section) to develop an account of that notion. The proposed model of 'implementation' provides a better understanding of the nature of levels of description, in particular for cognitive mechanisms. I argue that the standard trichotomy of levels of description, due to David Marr, is thus too coarse-grained. More importantly, the present analysis implies that the use of rational analyses is entirely orthogonal to the level of description for a theory. The level of description dictates a particular level of realist commitment to theoretical elements, while rational analysis is a tool by which to develop the theories. No close connection between rational analyses and a particular level of description is warranted, and so rational analyses can (and should) be used much more widely than they currently are.

Rational Analyses and Explanations

The fundamental idea behind a rational analysis is the intuition that people are behaviorally well-adapted to their environments (Anderson, 1990, 1991a; Chater & Oaksford, 2000; Oaksford & Chater, 1998). More specifically, rational analyses understand 'rational' to be a relation that holds just when there is an appropriate connection between four complicated variables: (i) the relevant environment; (ii) the task or problem to be solved; (iii) the agent's capacities; and (iv) the agent's input and output (typically, but not necessarily, perception and behavior). In particular, 'rational' is taken to mean: "the agent's input/output are optimal[1] for solving the task in this environment, given the agent's capacities." A rational analysis consists of

[1] The cost function for evaluating optimality is typically specified as part of element (ii), the problem to be solved.

specifying any three of these elements, and then deriving the value for the fourth that completes the relation of rationality. In practice, the most common rational analysis specifies the first three elements and then derives the optimal input-output function. That is, one determines the optimal behavior to solve some task given the environment, input, and agent capacities, and then typically aims to confirm that model using various empirical data. Although less common, there are three other species of rational analysis that correspond to specifying a different subset of variables. For example, in order to determine why people exhibit some particular behavior, one might carry out a rational analysis in which the problem is the free variable: that is, one could determine the task for which some given behavior would be optimal (given the environment and agent). The type of rational analysis determines what factor is being explained, and which other three factors must be independently specified (see also Sloman & Fernbach, this volume).

Rational analyses aim to provide explanations for behavior, and they typically make no commitments about the underlying implementing mechanisms. One of the central benefits of a rational analysis is that—when done fully and correctly (see below)—it is capable of supporting a particular type of explanation about behavior: namely, one that answers questions such as "why is this behavior optimal?"[2] Optimality-based explanations give a reason for why the world has certain features, rather than simply asserting that those features exist, and are therefore often thought to provide superior, or more "normative," explanations than ones that do not appeal to optimality. If we can give an optimality-based explanation for some behavior, then the behavior is not simply an accident, and so, for example, we can infer that it would be likely to arise, even

[2] There is (almost) philosophical consensus that all explanations are answers to why-questions, but a long history of debate about exactly what constitutes a satisfactory explanation (Bromberger, 1993; Halonen & Hintikka, 2005; Hempel, 1965; Salmon, 1984, 1989; van Fraassen, 1980). Even for superficially simple why-questions, such as "why does behavior *X* occur?", there are many different types of responses that could provide an answer.

if the world were different in certain ways. Much of the discussion about optimality-based explanations—particularly in the philosophy of biology—has focused on exactly what standards must be met by claims that some behavior results *because* of its optimality (Brandon & Rausher, 1996; Orzack & Sober, 1994, 1996; Seger & Stubblefield, 1996). An optimality-based explanation is typically understood to require the defense of three related claims: (i) behavior *X* occurs in situation *Y*; (ii) *X* is optimal in situation *Y*; and (iii) people do *X* because it is optimal. The first and second components are closely connected, as a characterization of the situation is a necessary precursor for demonstration of optimality. The first element is a primarily descriptive task; in a cognitive science context, this corresponds to empirical evidence of a particular behavior pattern. The second part is primarily theoretical/mathematical; the mathematical work of a rational analysis corresponds to the necessary demonstration of optimality.

The third component of an optimality-based explanation is quite different from the other two: it requires one to show that the optimality of *X* played a causal role—typically, a major causal role—in people's doing *X*. One must show that *X*'s optimality produced some "pressure" that pushed people (in some sense) towards behavior *X*, and so one must specify a mechanism or mechanism-schema by which behavior is shaped and moved towards *X*, as well as historical evidence that the mechanism was active and relevant. The term 'cognitive mechanism' is notably absent from this description of an optimality-based explanation, since the explanation focuses on changes in the *behavior*, *rather* than in any mechanism for the proximal production of the behavior. One can give an optimality-based explanation without ever appealing to, or even describing, the underlying cognitive mechanism that produces the behavior. Of course, ignorance about the underlying mechanism might make it significantly harder to defend the third claim, but

descriptive knowledge of the cognitive mechanism producing *X* is not a necessary condition of an optimality-based explanation.

There are two standard types of stories that have been provided for the third component. Probably the most common use of optimality-based explanations is to provide explanations of biological traits as evolutionary adaptations. In that domain, the third component comes from a (partial) demonstration that some trait *A* is widespread in a population because of natural selection on precursors of *A* in previous generations. Selection pressures generally act to increase the proportion of more optimal traits in a population, and decrease the less optimal traits. Thus, if *X* is optimal (and evolutionarily available), then selection pressures can explain why behavior *X* emerges (rather than *A, B,* or…). Evolutionary explanations of *X*'s occurrence in terms of its optimality trace a causal sequence in phylogenetic (i.e., evolutionary) time. The other standard route to providing the third component is through an account in ontogenetic time (i.e., individual development) that shows how the optimal behavior is acquired through learning mechanisms. One can explain behavior *X* by showing that people learn (perhaps implicitly) that *X* is the optimal behavior in some domain (and they do *X* for that reason). In both types of accounts, the primary focus is on the optimality of behavior in "normal" situations, since those are the ones in which the organism or lineage actually encounters pressure to behave optimally, whether in phylogenetic or ontogenetic time. Optimality-based explanations are thus based principally on ecological norms (Gigerenzer, 2000, 2006; Gigerenzer & Selten, 2001): what matters most is the fit of a behavior with the situation and environment, not some abstract theoretical ideal.

There are tradeoffs between the two routes. For phylogenetic stories, there is no difficulty specifying or establishing the appropriate mechanism providing "pressure" towards optimality; natural selection requires no defense. One does, however, have to show that natural selection was

causally relevant for this trait, and there are significant measurement challenges for such demonstrations, though they can sometimes be overcome (as in, e.g., Gilchrist, Azevedo, Partridge, & O'Higgins, 2000; Losos, 1992; McCracken, Harshman, McClellan, & Afton, 1999). In contrast, an account in ontogenetic time faces many fewer measurement challenges, since one can track the changes over time in an individual's behavior (or the behavior of a population). That story must, however, show that the behavioral changes were the result of some "pressure towards optimality," and it is frequently much harder to demonstrate that *optimality* was the reason for change. These tradeoffs naturally suggest a two-stage account: namely, provide an ontogenetic account for the occurrence of behavior *X* in the individual's lifetime, and then argue that the necessary "pressure towards optimality" arises from a learning mechanism that has been optimized in phylogenetic time to find the optimal behavior in any particular situation. Regardless of the details, however, some story must be provided to have a full optimality-based explanation; one cannot simply stop with the assertion that "behavior *X* occurs and is optimal," as it does not follow that optimality was the reason for *X*'s occurrence.

Rational analyses aim to provide optimality-based explanations: "a rational explanation tells why the mind does what it does" (Anderson, 1991b, p. 410). In practice, however, rational analyses almost always consist solely in optimality analyses that show that a particular behavior is optimal for an environment, agent, and problem, followed by experiments to confirm that the optimal behavior occurs. An optimality analysis alone, though, is insufficient for an optimality-based explanation, as there are many other reasons why *X* might occur. People might act optimally because of historical accident, or because there are no other options, or a number of other reasons. One can use an optimality argument as a heuristic device to suggest a novel descriptive theory (as suggested by Sloman & Fernbach, this volume), but an optimality-based

explanation requires more. Rational analyses that fail to provide an account of how the behavior came to be fall short of full optimality-based explanations. There is no principled barrier to providing such explanations; rational analyses just have not typically done so in practice.

Mechanism-based explanations offer a very different type of explanation than optimality-based explanations. They aim roughly to describe the proximal sequence of causes that led to the behavior: they provide an answer to "how behavior *X*?" It is easiest to think about mechanism-based explanations in terms of an over-simplified picture in which behavior results from cognitive mechanisms proceeding from some initial conditions. This picture is of course highly unrealistic, as it ignores the dynamic nature of cognition, the difficulty of individuating cognitive mechanisms, and so on. However, given this picture of the proximal causes of behavior, we can understand a mechanism-based explanation as one that provides a causal sequence that starts with initial conditions, and then moves through the cognitive mechanisms to a final effect that is the behavior *X*. Mechanism-based explanations are thus entirely descriptive: they point towards cause-effect sequences in the world without an attempt to explain why that particular sequence is the one that occurs. We can offer such explanations to account for both token instances of behavior *X* by a description of specifics of a particular situation, and type-level descriptions of behavior *X* by description of "standard" conditions and mechanisms. One frequently provides only the initial conditions or mechanism in a mechanism-based explanation for pragmatic reasons: one assumes that the other element is widely known. Most mechanism-based explanations offered in psychological papers, for example, focus on cognitive mechanisms rather than initial conditions, since the latter are usually specified in the description of the experiment.

Rational analyses essentially never offer mechanism-based explanations. Much of the literature on rational analysis is quite explicit with regards to its agnosticism about the

underlying cognitive mechanisms (e.g., Anderson, 1990, 1991b; Chater, Oaksford, Nakisa, & Redington, 2003; Tenenbaum, Griffiths, & Niyogi, in press). A nice expression of the attitude is Anderson's (1991a) statement that rational analyses should "prescribe what the behavior of a system should be rather than how to compute it" (p. 483). This basic attitude is also captured by descriptions of rational analyses as providing "as-if theories": the agent acts *as if* she were performing the computations of the rational analysis, but need not actually do them. The standard justification for this agnosticism is in terms of the influential framework of levels provided in Marr (1982). Rational analyses offer theories at the *computational level*: the level of description that focuses on goals and problems for an information processor, as well as input-output specifications of solutions.[3] By focusing on high-level characterizations of the problems and processes of an agent, proponents of rational analyses aims to sidestep questions of underlying mechanism or architecture since those are (the argument runs) not directly relevant to questions and concerns at the computational level. Any implementation that computes the particular (optimal, rational) input-output function is acceptable, and so cognitive science can proceed by focusing simply on that level.[4] In practice, of course, rational analyses sometimes have some

[3] A different view (suggested to me by Josh Tenenbaum) is that Marr's levels provide the scope of applicability: the computational level centers on problems that any agent must solve, the algorithmic level on humans, and the implementation level on neuroscience. Marr-levels in this interpretation do not correspond to "level of description," though the computational level will still almost always be focused on input-output, since the underlying mechanism is not directly relevant at that level.

[4] Some proponents of rational analyses seem to be motivated by an additional worry: if one can only ever measure behavior (and external situations), then one cannot ever uniquely determine the underlying cognitive mechanisms, and so input-output is the best we can do. For example, Anderson (1990) writes: "behaviorism was correct…that you cannot infer the mechanisms in the black box from what goes in and what goes out" (p. 26). But although reasoning about unobserved entities and processes is hard, it is certainly not impossible. It is absolutely standard in the natural sciences, and no argument has been provided that cognitive psychology is somehow special in this regard.

connection with underlying mechanisms, but the ideal is of strong separation between the rational analysis and implementing mechanism.

Rational Analyses and Instrumentalism

The strong agnosticism of rational analyses about cognitive mechanisms means that rational analyses are almost always instrumentalist: no commitment is made to the physical or mechanistic reality of the internal structure and computations of the model; they need not correspond to anything in the world. The mathematics is purely a computational device by which one can generate (hopefully) accurate predictions. If placed in an appropriate optimality-based explanation, then a rational analysis can provide more than a simple description, as we can say why the input-output function has the structure it does. The explanation is (when available) only at the level of structure in the <input, output> pairs, but at least it is *some* level of explanation. This observation serves to highlight the importance of the oft-neglected developmental component of an optimality-based explanation. Without it, rational analyses are strongly instrumentalist descriptions of an individual's stimulus-response function, coupled with the observation that the function is optimal (though we do not know if that optimality is the reason for the individual's actions). Instrumentalist theories are also unable, even in principle, to provide mechanism-based explanations, since they avoid any discussion of the underlying cognitive mechanisms.

The lack of anchor in a mechanistic account points towards at least two conditions when one should have much less confidence in the predictions of an instrumentalist theory. The first situation is when one aims to use the instrumentalist theory outside of the domains for which it was originally developed. The claim here is *not* that instrumentalist theories can never be used in novel domains; rather, the concern is that it is much harder to determine for an instrumentalist

theory whether it has a sufficiently broad foundation of empirical data to generalize successfully to novel domains. In general, one can transfer theories to novel cases when the underlying causal structures of the two situations are suitably similar. This standard translates in practice into having sufficient coverage in the empirical data to know the approximate causal or correlational structure (as appropriate) in the various conditions. If one has no knowledge of the underlying mechanisms, then there are no natural standards by which to judge whether the confirming empirical data derives from a suitably broad set of conditions. Instead, one simply has to collect data from a wide range of conditions—presumably informed by experiments showing which factors seem to be relevant—and then argue that this set provides sufficient coverage. There are obviously methods to mitigate this problem and ensure suitable generality for one's theory, but the general problem is significantly worse for an instrumentalist theory than for a mechanistic one. Moreover, in the case of rational analyses, we have positive reasons to think that the behavior *will* differ significantly in unusual cases. Rational analyses aim to understand behavior as optimal in "standard" environments, and if behavior is actually tuned to the environment (as argued by many proponents of rational analyses), then we should expect behavior in unusual cases to be quite *sub*-optimal, and so not conform to a rational analysis that incorporates that situation.[5]

In one particular case, mechanism information is actually necessary for prediction. In cognitive science, we often need to predict or explain an individual's behavior when the system "breaks" in various ways, whether because of external trauma, various mental deficits, or perhaps even local features of a situation (e.g., extreme time pressure). As the literature on

[5] This is another point of agreement between proponents of rational analyses and of ecological rationality: both think that "failures" by experimental participants are often because the experiment fails to match the structure of the everyday environment along important dimensions.

causation repeatedly reminds us (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993), predictions about the effects of interventions, manipulations, or changes in the system require knowledge of the underlying causal structure. Instrumentalist theories provide no information about mechanisms, and so no grounds for predictions when the cognitive system breaks. In the particular case of a rational analysis, there is no reason to think that a broken cognitive system will behave optimally, or even close-to-optimally. We cannot just plug the boundary conditions/initial stimuli into the rational analysis to predict the behavior of a broken cognitive system, since we require information about how the system works. As an analogy, if we want to predict what a calculator will do (if anything) if we hit it with a hammer, then knowledge of the input-output functions will not suffice. We need to know how the underlying circuitry works to make accurate predictions.

A more foundational problem is that the very goal of true agnosticism about mechanisms is ultimately a will-o'-the-wisp: even the most diehard proponent of agnostic rational analysis must eventually talk about underlying cognitive mechanisms in various ways. Rational analyses, for example, are supposed to incorporate computational limits, but the nature of those limits depends partly on the underlying cognitive architecture. Putatively instrumentalist rational analyses must have some connection to the underlying mechanisms, or possibility space for the mechanisms. One should not view this dependence as some impurity that should be purged; in general, connections between one's various cognitive theories are clearly a virtue. Some degree of instrumentalism is acceptable and even necessary (see the remainder of this section), but the agnosticism of rational analyses unnecessarily (see section 4) forces a pervasive instrumentalism that denies the possibility of rational analyses providing mechanistic explanations.

There are positive aspects to treating rational analyses as instrumentalist theories. Instrumentalism frequently has a bad name in the philosophy of science, but all theories—even putatively mechanistic ones—have some degree of instrumentalism, in that certain cognitive operations are left unexplained. For example, associationist models of causal learning (or more generally, stepwise error-correction models) are often characterized as mechanistic theories, but they provide no account of how the prediction is generated, how the error is calculated, and so on. Those theories provide a mathematical account of *what* is calculated (including various intermediate representations, such as associative strengths, error, and so on), but no explanation of *how* those intermediate calculations are carried out.[6] Associationist models—like all theories—have components that are treated instrumentally.

Completely instrumentalist theories can also be appropriate and useful in certain conditions. Instrumentalist theories provide accurate descriptions of the data, as well as some predictions for novel situations. Well-confirmed instrumentalist theories at least have the virtue of descriptive accuracy and predictive power for situations in their purview. Although it is tempting to discount the worth of descriptive adequacy, numerous episodes in the history of science reveal how hard it can be to find a theory that captures the observed data, even if we set aside issues of realism. We should be careful not to downplay the importance of finding a descriptive theory, particularly if it also provides reasonably accurate predictions. These descriptions can sometimes be computationally simpler than mechanistic models of the same data, since the instrumentalist theories are not bound by fidelity to the underlying mechanisms.

Accurate descriptions of input-output functions also enable us to characterize the relevant factors for some cognitive problem, even if we do not know precisely how that factor is used in

[6] There are of course various neural accounts of associative learning (e.g., Maren, 2001; Menzel & Giurfa, 2001), but those are additional theories.

the cognitive system. For example, rational analyses of causal reasoning highlight the importance of base rate information when one has ambiguous data, and that importance has been experimentally confirmed (e.g., Sobel, Tenenbaum, & Gopnik, 2004), though we do not know the mechanism by which base rate information is incorporated. Similar stories are available for other domains: rational analyses suggest that certain, previously ignored, environment and task features are cognitively relevant, and subsequent experimental research confirms the importance of those features. Instrumentalist theories are better able to highlight the relevant features of a situation, precisely because they are not committed to a mechanism. Overarching tendencies and generalizations can be straightforwardly expressed in the mathematics of the model.

Complete theories certainly might be the desired end-state, but we need to recognize that much of science occurs in domains that we understand only poorly. Instrumentalist theories are metaphysically less ambitious, and so face a lower standard for confirmation and acceptance. They are more insulated from criticism since they can only be critiqued for failure to match the observed data. As a purely practical matter, one should sometimes pursue an only instrumentalist theory, with the understanding that a mechanistic theory will follow. In particular, if one wants to argue that some previously ignored factor is cognitively relevant, then instrumentalism allows one (by virtue of its metaphysical caution) to focus on that factor, rather than also arguing for a particular account of the underlying cognitive mechanisms. Rational analyses can be particularly helpful in these domains, as the assumptions of optimality provide a defeasible guide to the development of a novel theory. There is at least one descriptive theory—namely, the theory that says people act optimally for this task, environment, etc.—that is salient in the search space. Rather than searching quite broadly through the descriptive theory-space, one can target the optimal and near-optimal models for initial testing.

Rational Analyses and Implementations

The theses of the previous two sections can be roughly summed up as: rational analyses are almost always instrumentalist (section 1), and instrumentalist theories are generally undesirable, though they can be useful in some situations (section 2). One might thus make the natural inference that rational analyses are generally undesirable, but this inference presupposes that the first premise follows from a necessary feature of rational analyses. I claim, however, that the close tie between rational analyses and instrumentalist theories is unwarranted. The instrumentalism of rational analyses derives largely from the connection with the computational level of Marr's trichotomy, and this section argues in two stages that this connection is neither necessary nor desirable. First, I argue that a careful consideration of the notion of 'implementation' shows that (a naïve reading of) Marr's trichotomy of levels is overly coarse, and forces instrumentalism where none is warranted. Second, and more importantly, this account of 'implementation' makes clear that the use of rational analyses is entirely orthogonal to the particular level of description. Proponents of rational analyses have unnecessarily tied themselves to thinking about behavior and cognition at one level of description. Rational analyses can (with appropriate care) have a much wider range of application.

The first question is what it means for some mechanism *M* to implement a rational analysis *R*. The standard response in the rational analysis (and related) community seems to be: *M* implements *R* whenever the target behavior for which *M* provides a mechanism-based explanation is the same as the behavior for which *R* provides an optimality-based explanation. In practice, this means that *M* implements *R* whenever the input-output function for mechanism *M* is approximately the same as the optimal function described by *R*. The inputs and outputs are assumed to be (something like) stimuli and behavior, respectively; all that matters is same

behavior in same situation. This understanding of 'implementation' fits quite nicely with the standard use of rational analysis at the computational level. As just one example, Take-the-Best, combined with the Recognition Heuristic, provides a mechanism explanation of behavior in binary choice problems with limited information (Gigerenzer, 2000; Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 1999). Except in special situations (e.g., non-compensatory cues), Take-the-Best is not a "rational" algorithm in the sense of being optimal over all cases, as it sometimes ignores information that is potentially useful in decision-making. Nonetheless, Take-the-Best has been described as "rational" because it produces behavior that closely approximates the optimal behavior in this context (Chater, *et al.*, 2003; Gigerenzer, Czerlinski, & Martignon, 1999; Martignon & Hoffrage, 1999). That is, the argument runs: Take-the-Best plausibly implements (the model in) a rational analysis because it produces appropriate behavior, and not because of any deeper connection or similarity.

The proposal that 'implementation' means "input-output function approximation" appeals to the intuition that behavior is ultimately all that really matters; correct behavior is the only means for characterizing a system. This notion does seem to capture an important aspect of implementation: namely, that we can have different implementations of the same high-level algorithm. It also seems too weak to be a full characterization of implementation. For example, we may want to require an implementation of rational decision-making to have separable intermediate representations of utility and probability, but the current definition prohibits us from imposing any such restriction. An implementation (in the input-output sense) of a rational analysis also cannot provide any optimality-based explanation beyond what was already available in the rational analysis (Danks, submitted). This relation of implementation leads to additional understanding of how behavior is generated, but no additional understanding about

why (in the optimality sense) the behavior occurs. The intertheoretic relation is so weak that the theories standing in it provide neither constraints nor information for one another.

Philosophical work on intertheoretic relations suggests a different understanding of 'implementation': namely, that the mathematics of the high-level theory is a special or limiting case of the mathematics of the lower-level theory (Batterman, 2002; Nickles, 1973; Rueger, 2001, 2005; Smith, 1998). For example, Newtonian mechanics in a collection of indistinguishable particles is a limiting case implementation of the ideal gas law, since the latter is a limiting case of the relevant statistical mechanics equations in the former. This understanding of implementation can also hold between theories at the same level of description: Newtonian mechanics is a limiting case of relativistic mechanics as $(v / c)^2 \rightarrow 0$. This notion of limiting case implementation of a rational analysis is applicable to some psychological theories: for example, Griffiths & Tenenbaum (2005) argue that a mechanistic theory based on the χ^2 statistic is a limiting case implementation of rational causal learning using causal support (though probably not descriptively accurate). Much of the recent philosophical literature has focused on the difficulties that can arise when the limit process yields a singular limit (e.g., Batterman, 2002; Rueger, 2005); those issues have not arisen in the psychological literature.

Implementation *qua* limiting case relation is obviously a much stronger intertheoretic relation than input-output approximation, since it arises when the appropriate mathematical relation holds between the *equations* of the two theories, rather than just the input-output functions. The two notions of implementation are logically independent (Danks, submitted), though in practice, limiting case implementation almost always implies input-output implementation, but not *vice versa*. This additional strength carries with it certain benefits, as numerous theoretical virtues are arguably transferred between theories that stand in this relation

(Batterman, 2002; Nickles, 1973; Rueger, 2001). However, limiting case implementation seems to be too strong to use as a general understanding of 'implementation,' since it requires the intertheoretic connection to extend all the way down to the fundamental equations of the theories. This notion does not allow for some steps to be encapsulated in a "black box." An appropriate massive look-up table seems like it should count as a type of implementation of addition, even though the algorithm in the look-up table (e.g., a hash code) is not a limiting case of "real" addition.

I suggest instead that 'implementation' is better understood by thinking about computer programming. In general, we can think about a computer program as a series of function or method calls, where all functions are defined in terms of either (i) a different series of function calls; or (ii) a particular input-output relation. Much of the literature on good programming techniques (e.g., Lakos, 1996; McConnell, 2004) focuses on exactly the question of how to divide up the program into suitable components. I propose then that a program *L* implements a program *U* just when we can transform *L* into *U* by a set of computable transformations, each of which takes (a) some sub-series of function calls in *L* and (b) replaces it with a single function call from *U* with the same input-output relation as the sub-series. That is, if we can "bundle" sequences of functions in *L* to yield *U*, then we say that *L* implements *U*. I contend that most rational analyses—and more generally, essentially all cognitive theories that make quantitative predictions—can be thought of as programs (broadly construed). This notion of 'implementation' thus transfers straightforwardly to the question of when a mechanistic theory *M* implements a rational analysis *R*.

I do not intend anything particularly fancy by this characterization of 'implementation.' Similar ideas about the nature of implementation and hierarchies of description can be found in

box-and-arrow diagrams in cognitive neuropsychology, or in the widespread use of flowcharts or dataflow diagrams to describe cognitive theories. The central difference here is that computer programs—and so my account of 'implementation'—typically require much more specificity about the functions than one finds in standard box-and-arrow diagrams. A dataflow diagram can have a box labeled "Pragmatic Considerations" without any precise characterization of the input-output function for that box; the same function call in a program or fully-specified theory would have to specify those input-output relations.

On this characterization of 'implementation', input-output approximation corresponds to the case in which *U*—the theory being implemented—consists of a single function call that is defined by an input-output relation. Thus, any sequence of other function calls that has the same input-output relation will implement *U*. Limiting-case approximation is roughly the situation where *U* and *L* are nearly identical in terms of close correspondence between both function calls and the internal sequence, but *L* is slightly more specific, or uses a slightly different input vector. If the two programs are identical except that *L* uses a slightly different—and usually more general—method at a particular place in the computation, or *L* uses slightly more information, then *L* is a limiting case implementation (given a few technical constraints). Both types of implementation are therefore special cases of this characterization of 'implementation.' In general, the granularity of specification in *U* sets an upper bound for all implementing theories: *L* cannot implement *U* unless its sequence of function calls is more specific.

This notion of 'implementation' appropriately captures the earlier observation that all theories are partially instrumentalist, and none are completely instrumentalist. The ubiquity of instrumentalism arises from the fact that all cognitive theories have some "basic" (relative to the theory) operations or functions for which no mechanism is provided, just as all computer

programs have function calls (e.g., addition) that are specified only by input-output relations. One cannot specify every function call in terms of other function calls without regress. At the same time, all computer programs must be sensitive to the function calls that are available in the architecture, and so must be aware of characteristics of the system on which the program is running. One cannot write or run a computer program in *complete* ignorance of the underlying hardware system,[7] and even the most agnostic rational analysis assumes that certain functions can somehow be computed.

Most importantly, this characterization of 'implementation' leads to a more refined understanding of levels of description. This notion implies that "level of description for a theory" really means something like "granularity of the realist commitments of a theory." By expressing a theory as a computer program, one must clearly identify which components of the theory should be interpreted realistically (i.e., the ordered sequence, and the function calls), and which only instrumentally (i.e., the internal computations in function calls specified only by input-output relations). One is committed only to realism about the stated function calls in the "cognitive computer program," and not to any particular way of generating that input-output relation. There is no ambiguity about what theories could implement this particular cognitive model. The complexity of interesting cognitive theories also suggests that there will very rarely be only three levels of description. One will typically be able to continue to "push downwards" to find even finer grains of description, and so there is no reason to accept Marr's trichotomy of levels (or at least, a naïve understanding of it).[8] There will typically be (almost) infinitely many

[7] Even for putatively cross-platform languages such as Java, important hardware constraints can periodically trickle up. Spolsky (2004, ch. 26) gives this phenomenon the wonderful title of The Law of Leaky Abstractions.

[8] The observation that Marr's levels are too coarse is certainly not novel to me. For example, Anderson (1990, p. 17) argues that we should use four levels.

levels of description, corresponding to the different ways that the function calls can be instantiated in sequences of other function calls, and so there is no sharp line distinguishing the computational, algorithmic, and implementation levels. Instead, there are many different degrees of realist commitments corresponding to the different granularities of specification of the cognitive program, ranging from a single function call at the most abstract, downward in a series of increasingly precise expressions of the program.

The current view in cognitive science that there is a special computational level is reinforced by the persistent identification of that level (i.e., a single function call) with rational analyses. The preceding discussion shows that the use of rational analyses *qua* optimality analyses is entirely orthogonal to the level of description. A rational analysis viewed through the present lens simply aims to find the optimal sequence of function calls for some task and environment *given some set of basic functions.* Most rational analyses use a set of basic function calls corresponding to (roughly) all computable functions and so have no need to consider sequences of function calls, but one can ask about the optimal sequence even when we substantially restrict the basic functions. The problem is made even more interesting if one imposes time constraints on the overall problem and time costs for each of the function calls.[9] In general, the optimality of a program is determined by its performance relative to its competitors; the level of description is irrelevant, except when the level is used to define the possibility space. Rational analyses thus actually have a much wider range of application that one typically finds. One can ask about the optimality of a particular mechanistic theory relative to other potential mechanistic theories. One can even determine the optimality of a theory relative to ones that have exactly the same overall input-output relation. The current agnosticism about mechanisms

[9] Without such a restriction, the space of programs (i.e., series of function calls) will typically cover all or almost all computable functions, even for quite restricted basis functions.

that one finds in rational analyses is unnecessary: a rational analysis can focus directly on the optimal mechanism for a particular problem, and not just the optimal input-output function. Rational analyses focus on a property of a theory (relative to its competitors) that does not depend on the level of description.

There is a pro-computational level argument that I have not yet considered. An optimality analysis aims to identify the highest-performing (by some metric) function from the search space. The explanatory power of the optimality analysis is arguably positively correlated with the size of the search space: if the optimal function is determined more by the lack of serious competitors than by high performance, then the *optimality* of the function seems to provide little explanatory power. The lack of competition seems to be explaining the occurrence of the behavior. Thus, a rational analysis is thought to be more powerful if the observed behavior is optimal relative to a large set of competitors, and the relatively unrestricted computational level provides the largest such set.

This argument is superficially appealing, but I think that it makes a major mistake. Consider the general claim about the correlation between the explanatory power of an optimality analysis and the size of the search space. The claim of a correlation is plausible: if the search space is small, then it is (arguably) less likely that the optimality of the function is the reason for the individual having that function, and so it is correspondingly less likely that there is any optimality-based explanation available for the behavior. (Recall that an optimality-based explanation requires an account of the development of the function in which the optimality of the function plays an important causal role.) But we are not simply interested in optimality analyses; we want optimality analyses—and optimality-based explanations—that are actually true of the world. A demonstration that some function is optimal relative to a large search space is

mathematically interesting, but it is not clear how it is relevant from a cognitive science point-of-view unless that optimality, relative to *that* set of competitors, played some role in the development or maintenance of the behavior. To establish this latter point, the search space for the optimality analysis needs to be approximately the same as the search space for the developmental story. One needs to use the *actual* possibility space, and not one that that would have provided more explanatory power, if only it had been the actual possibility space.

There is one final, practical argument in favor of research at the computational level. Search for the optimal function can be quite difficult, and is arguably made even more challenging if the possibility space is "complex" in various ways. The computational level—i.e., descriptions in terms of a single input-output function call—using all computable functions is a large search space with relatively simple structure. Thus (the argument continues), one should do optimality analyses at the computational level for practical reasons. Of course, to the extent that one subsequently learns more about the underlying physical and architectural constraints, one could amend the optimality analysis. But (the argument concludes) given our current (relative) ignorance about underlying cognitive architectures and implementations, one should do rational analyses at the highest level of description (unless one has substantial additional knowledge as in, e.g., Lee & Mumford, 2003).

This practical argument suffers from the same flaw as the previous argument: our rational analyses only can figure in optimality-based explanations if they use approximately the actual space of competitors. Mathematical tractability may be an important pragmatic reason for working at the computational level at first, but it does not justify an exclusive focus on that level. Moreover, it is not even obvious that the search space of single input-output function calls is dramatically simpler once one actually has some knowledge of the possibility space of

underlying mechanisms. It is often much easier to specify the space of possible functions in terms of the intermediate steps, rather than remaining at solely the input-output level. Any particular *program* will be described more simply by a single function call, but the *space* of possible programs will not necessarily be described more simply in those terms. It just is not the case that the description of the function possibility space at the level of single input-output function calls is always simpler than a description in terms of intermediate function calls. Thus, the claim that optimality analyses at the level of single input-output function calls will always be more mathematically tractable falls apart. Much depends on how much information one has about the underlying architectures or possible implementations.[10]

Conclusion

Rational analyses have, without a doubt, been a useful tool in finding novel descriptive theories, previously unknown causally relevant task and environment variables, and have provided powerful frameworks for modeling complex cognition. I have nonetheless argued that their current use is not quite right. Extant rational analyses almost never provide full optimality-based explanations, as they almost never provide a developmental story—either phylogenetic or ontogenetic—that shows how the optimality played a major causal role in the establishment or maintenance of the behavior. By focusing on Marr's computational level, proponents of rational analyses have turned their attention away from theories of underlying cognitive mechanism, and the result has been rational analyses that provide essentially no explanations at all; they are just instrumentalist accounts. But as I have tried to argue in the last section, this tie to the computational level is entirely unnecessary: conditional on a set of competitors, the level of

[10] Of course, my reply assumes that human cognition is not arbitrarily plastic (in the relevant sense). If people can actually implement almost any computable function, then the restriction to the broadest search space at the computational level is probably justifiable.

description of a theory is irrelevant to its optimality. The level of description might shape the possibility space in certain ways, but that does not imply that rational analyses can only be carried out at the highest levels of description. Rational analyses can and should be performed at all levels of description, depending on the particular knowledge one has about the underlying cognitive mechanisms at that level.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*, 471-484.

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

Batterman, R. W. (2002). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford: Oxford University Press.

Brandon, R. N., & Rausher, M. D. (1996). Testing adaptationism: A comment on Orzack and Sober. *American Naturalist, 148*, 189-201.

Bromberger, S. (1993). *On what we know we don't know: Explanation, theory, linguistics, and how questions shape them*. Chicago: University of Chicago Press.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese, 122*, 93-131.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90*, 63-86.

Danks, D. (submitted). Explanatory power, input-output approximation, and the case of rational analysis. *Philosophy of Science*.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Gigerenzer, G. (2006). Bounded and rational. In R. J. Stainton (Ed.), *Contemporary debates in cognitive science* (pp. 115-133). Oxford: Blackwell.

Gigerenzer, G., Czerlinski, J., & Martignon, L. (1999). How good are fast and frugal heuristics? In J. Shanteau, B. Mellers & D. Schum (Eds.), *Decision research from Bayesian approaches to normative systems* (pp. 81-103). Norwell, MA: Kluwer.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650-669.

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox.* Cambridge, MA: The MIT Press.

Gilchrist, A. S., Azevedo, R. B. R., Partridge, L., & O'Higgins, P. (2000). Adaptation and constraint in the evolution of Drosophila melanogaster wing shape. *Evolution & Development, 2,* 114-124.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 37-58). New York: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Halonen, I., & Hintikka, J. (2005). Toward a theory of the process of explanation. *Synthese, 143,* 5-61.

Hempel, C. (1965). *Aspects of scientific explanation*. New York: Free Press.

Lakos, J. (1996). *Large-scale C++ software design*: Addison-Wesley Professional.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20,* 1434-1448.

Losos, J. B. (1992). The evolution of convergent structure in caribbean Anolis communities. *Systematic Biology, 41,* 403-420.

Maren, S. (2001). Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience, 24*, 897-931.

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

Martignon, L., & Hoffrage, U. (1999). Where and why is 'take the best' fast, frugal and fit? In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 119-140). Oxford: Oxford University Press.

McConnell, S. C. (2004). *Code complete* (2nd ed.). Redmond, WA: Microsoft Press.

McCracken, K. G., Harshman, J., McClellan, D. A., & Afton, A. D. (1999). Data set incongruence and correlated character evolution: An example of functional convergence in the hind-limbs of stifftail diving ducks. *Systematic Biology, 48*, 683-714.

Menzel, R., & Giurfa, M. (2001). Cognitive architecture of a mini-brain: The honeybee. *Trends in Cognitive Sciences, 5*, 62-71.

Nickles, T. (1973). Two concepts of intertheoretic reduction. *The Journal of Philosophy, 70*, 181-201.

Oaksford, M., & Chater, N. (1998). *Rational models of cognition*. New York: Oxford University Press.

Orzack, S. H., & Sober, E. (1994). Optimality models and the test of adaptationism. *American Naturalist, 143*, 361-380.

Orzack, S. H., & Sober, E. (1996). How to formulate and test adaptationism. *American Naturalist, 148*, 202-210.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rueger, A. (2001). Explanations at multiple levels. *Minds and Machines, 11*, 503-520.

Rueger, A. (2005). Perspectival models and theory unification. *British Journal for the Philosophy of Science, 56*, 579-594.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Salmon, W. C. (1989). *Four decades of scientific explanation*. Minneapolis, MN: University of Minnesota Press.

Seger, J., & Stubblefield, J. W. (1996). Optimization and adaptation. In M. R. Rose & G. V. Lauder (Eds.), *Adaptation* (pp. 93-123). San Diego: Academic Press.

Sloman, S., & Fernbach, P. M. (this volume). The value of rational analysis: An assessment of causal reasoning and learning. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition.* Oxford: Oxford University Press.

Smith, P. (1998). Approximate truth and dynamical theories. *British Journal for the Philosophy of Science, 49*, 253-277.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*, 303-333.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Berlin: Springer-Verlag.

Spolsky, J. (2004). *Joel on software: And on diverse and occasionally related matters that will prove of interest to software developers, designers, and managers, and to those who, whether by good fortune or ill luck, work with them in some capacity*. New York: Springer-Verlag.

Strevens, M. (2003). *Bigger than chaos: Understanding complexity through probability*. Cambridge, MA: Harvard University Press.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (in press). Intuitive theories as grammars for causal inference. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.