Samuli Pöyhönen
(University of Helsinki)

Sidestepping complexity:
Rational analysis explanations in cognitive science

In cognitive science, In cognitive science, disagreements concerning the explanatory value of new probabilistic models of cognition have given rise to a strongly polarized debate (Chater & Oaksford 2008; Glymour 2011; Jones & Love 2011). Proponents of the rational analysis (RA) approach to cognitive modeling claim that their models can produce understanding of the nature of human cognitive capacities without relying on any mechanistic knowledge of cognitive processes or representations. This way of sidestepping the evidential and explanatory problems arising from the causal complexity of cognition is in deep tension with the influential causal-mechanistic view of scientific explanation, according to which genuine explanations must describe actual causal structures. In this paper, I disentangle various different explanatory contributions that have been attributed to RA models, and I assess the plausibility of such explanatory claims.

The rational analysis approach has recently become one of the prominent theoretical traditions in cognitive science (cf. Kousta 2010). The sophisticated modeling methods employed by the RA theorists appear to shed light on complex aspects of human cognition, which have often thought of as being beyond the reach of mechanistic research methods. Without making any substantial commitments about the underlying cognitive mechanisms, proponents of the RA approach have put forward novel theories of cognitive capacities such as memory, categorization, causal learning, and conditional inference (cf. Chater & Oaksford 2008). Often the resulting models fit empirical data far better than competing accounts, and the novel analyses of cognitive capacities provided by the models appear to have shed light on the nature of the explananda under study.

However, there is a large consensus in the philosophy of science that explanations also in the cognitive sciences should track causal mechanisms. In particular, as Kaplan & Craver (2011) argue, mathematical models in the mind sciences are genuinely explanatory only if there is a mapping between elements in the model and elements in the mechanism for the

phenomenon. Without such causal anchoring, models should not be conceived of as explanations, but rather as mere redescriptions of phenomena. This mechanistic constraint compromises the explanatory status of non-mechanistic RA models, and creates the puzzle of how to interpret the nature of their theoretical contribution.

Rational analysis modeling in cognitive science originates in John Anderson's (1990) work on memory and categorization. Relying on Marr (1982), Anderson pointed out that, at the time, mechanistic research in psychology was plagued by identifiability problems. Lacking a clear picture of what it is that cognitive mechanisms do (i.e. of psychological explananda), evidence of neural and algebraic level structures was insufficient to uncover the mechanistic architecture of the human mind. Much of the contemporary theorizing in the RA tradition has followed Anderson's drastic methodological move that tries to avoid the evidential and explanatory problems of mechanistic research by building models that are agnostic about all algorithmic and implementation level details. Instead, RA relies on three basic starting points: (1) Cognition is understood as probabilistic (Bayesian) computation. (2) The likelihoods and priors required by the probabilistic model are acquired from the analysis of environment structure. (3) Behavior of human agents is assumed constitute an optimal response to the task. Consequently, behavioral data can be used to calibrate or evaluate the models.

I illustrate this theoretical perspective by discussing Oaksford & Chater's (1994, 2007) model of the Wason selection task. Their information-gain model challenges the traditional logic-based theories of the nature of subjects' behavior in the task. They argue that behavior typically seen as irrational can be understood as an optimal way of decreasing uncertainty regarding the hypothesis studied.

Now, how (or whether at all) do such models provide explanations of the studied cognitive capacities? By relying on the contrastive-counterfactual theory of explanation (Woodward 2003), and the inferential account of explanatory understanding (Ylikoski & Kuorikoski 2010), I analyze three different kinds of explanatory contributions such models could be seen to make:

(1) Clearly, RA models do not allow any what-if inferences typical of constitutive mechanistic explanations, which relate changes in components of the system and their organization to the

properties of the macro-level explanandum (i.e. the cognitive capacity). While the theoretical commitments of RA modelers vary, typically the variables in the model are not given any psychological interpretation, and hence many modelers would themselves agree with this conclusion.

(2) Instead, it is a central tenet of RA that the models track explanatory dependencies between changes in the environment structure and corresponding changes in the behavior of the agents. As critics have pointed out, there are problems with such explanatory claims as well (Jones & Love 2011; Marcus & Davis 2012). In particular, the environment-behavior inferences involved in such explanations rely on the strong optimality assumptions, and I argue that neither empirical data nor considerations from evolutionary theory provide support for the stability of counterfactual environment-behavior dependencies beyond the typical environment circumstances.

(3) However, there is also a third way in which RA models could be said to lead to increased understanding of explananda. Provided that the description of the task environment does not arise from mere speculation but solid empirical research (cf. Martignon & Hoffrage 1999), mathematical analysis of the environment together with knowledge of the cognitive constraints of agents can make possible the exploration of the *logic of the situation* – the possible space of action for cognitive agents. In other words, instead of as an explanation of actual behavior, a rational analysis of the task can be understood as an inferential aid that relates values of environment parameters to optimal behaviors across a range of different environment states.

This interpretation of RA models makes them models of environment rather than of cognitive systems. However, as emphasized by Simon (1990), and perhaps somewhat neglected by the mechanistic approaches of explanation, human behavior can be seen as shaped by a scissors whose blades are the structure of task environments and the computational capabilities of the actor. Hence, while RAs might not be explanatory of human cognition in themselves, they can lead to increased understanding by complementing mechanistic theories with precise models of the crucial environment factors within which cognitive mechanisms are embedded.

References:

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

Chater, N. & Oaksford, M. (eds.) (2008). The Probabilistic Mind. Prospects for Bayesian Cognitive Science. Oxford: Oxford University Press.

Glymour, C. (2011). Osiander's psychology. *Behavioral and Brain Sciences*, 34 199-200.

Jones, M. & Love, B. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *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.

Kousta, S. (2010). Approaches to cognitive modeling. *Trends in Cognitive Sciences*, 14.

Marcus, G. & Davis, E. (2012). How robust are probabilistic models of higher-level cognition? *Psychological Science*, OnlineFirst.

Marr, D. (1982/2010) Vision: A computational investigation into the human representation and processing of visual information. W.H. Freeman/MIT Press.

Martignon L. & Hoffrage U. (1999). Why does one-reason decision making work? In Gigerenzer G., Todd P. M. & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 119–140). Oxford University Press, New York.

Oaksford, N. & 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: Oxford University Press.

Woodward, J. (2003). Making Things Happen: A theory of causal explanation. New York: Oxford University Press.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, 148, 201–219.