21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 481

Chapter 21

The value of rational analysis: an assessment of causal reasoning and learning

Steven Sloman, Philip M. Fernbach
Brown University, Providence, Rhode Island, USA

Our goal in this chapter is a rational analysis of human causal reasoning and learning. We take a rational analysis to be an assessment of the fit between data and a certain kind of model (Danks's chapter offers a more multi-faceted view of rational analysis). In the rational analysis tradition of Anderson (1990) and Oaksford and Chater (1998; in press), the term 'rational' has come to have three different meanings that vary in normative force. The first section of this chapter will be devoted to explicating these different meanings and evaluating their usefulness. The second section will apply these interpretations to assess the rationality of causal reasoning and learning.

The value of a rational model

In the rational analysis tradition, 'rational model' and 'computational model' tend to be used synonymously (e.g., Griffiths et al., in press). Danks (this volume) challenges this equation. According to Marr (1982), who introduced the computational level of description, a computational model describes the goal of a computation, why it is appropriate, and the logic of the strategy by which it can be carried out. What is missing from Marr's analysis is what determines the computation. Is it determined through an analysis of the task or must the analyst first observe what computation is actually being performed before engaging in a computational analysis? This is the critical question in determining whether or not a computation is 'rational.' Here are three different senses of 'rational model':

Normative model

This sense of rational model has its origins in Savage's (1972) analysis of subjective probability, a concept whose influence in psychology is primarily due to Kahneman and Tversky (1982). A rational model in this sense is a representation of the best way to perform a task. Given some goal, a normative model dictates what is necessary to achieve that goal. For instance, in the context of causal reasoning, if a machine is broken, a normative model might dictate the most cost-effective action to fix it.

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 482

AU: Please check whether the running head is ok.

The construction of normative models is a critical theoretical activity for two reasons. First, it is necessary for evaluating how well people perform. Normative models define optimal performance (relative to some goal) and in that sense set the standard for determining the validity of human judgment. Second, a normative model is necessary for isolating cognitive processes. On the assumption that people try to perform tasks well, people will always attempt to perform according to normative dictates. Therefore, good performance is a result of both cognitive processing and the constraints imposed by a task. But there is no way to distinguish the contribution of each and thus isolate the role of cognitive processing when performance is accurate. Only errors make the distinction possible because only errors do not reflect task constraints and thus must reflect cognitive processes. However, errors are defined in contrast to a normative model, a model of correct performance. In this sense, error provides a window onto processing, and the nature of processing can only be inferred in light of a normative model. Other aspects of processing, like time course or neural locus, don't necessarily require a normative model but often benefit from one.

The brunt of a normative analysis is to show that a particular type of model describes the best way to perform a task. That's the very object of a normative analysis; to show, for instance, that a Bayesian model of a task is normative for that task. In that sense, the modeling framework used to perform this kind of rational analysis is valid by hypothesis. It is not merely assumed, used for expository purposes, or used as a source of theoretical ideas. It is the object of evaluation. This distinguishes this form of rational analysis from the type espoused by Oaksford and Chater (in press), as we discuss below.

Normative analysis in the face of resource constraints

The type of normative analysis inspired by Simon (1955; 1956) evaluates performance not only with respect to the constraints imposed by a task – how best to perform a task – but also by the constraints that come to the task with the performer (cf. Cherniak, 1986; Harman, 1995). The performer has only limited time and energy, limited working memory capacity, limited knowledge, etc. Performance can be evaluated with respect to a model that describes optimal performance given a set of a priori constraints of this kind.

As long as the constraints are a priori and not chosen to make performance look more rational after the fact, this kind of model inherits the virtues of the simpler kind of normative model (i.e., type i). It defines a reasonable notion of rationality: doing the best that one is able to in order to achieve a goal. Moreover, it defines a model that supports inferences about the nature of processing. To the extent that a person deviates from the dictates of this model, it must be due to the nature of his or her mental processes and not to the constraints under which her or she is performing. In fact, it supports stronger inferences than the first type of model because errors with respect to this type of model cannot be attributed to the cognitive constraints that the model already embodies.

The costs of this kind of model are of two kinds. First, it is more complicated than the first kind because it must represent additional constraints. Second, its validity is more tenuous because it rests on more assumptions, namely that the constraints it imposes are real.

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 483

A model of what people are computing

The third sense of rational analysis, introduced by Anderson (1990) and further developed by Oaksford and Chater (in press), involves incorporating performance data into a normative modeling framework in order to use that framework to describe what people are actually computing. Generally this involves first choosing a modeling framework (normally Bayesian analysis) and then representing people's task assumptions within that framework. This kind of analysis involves a modeling plus testing cycle: A model is constructed out of a normative framework, its fit to data is assessed, then the model is changed to accommodate deviations, again its fit is assessed, and so on until an empirically adequate model is arrived at.

The model that results from this process does not have a claim to optimality or to normativity in any other sense because it is evaluated empirically, not in terms of the goodness of the actions that can be derived from it (Danks, this volume, argues that both are necessary). Hence this modeling enterprise is essentially descriptive, not rational. Once descriptive assumptions are incorporated, if those assumptions do not themselves have normative justifications, then any model that incorporates them has no claim to rationality. If the assumptions do have normative justification, then they should have been incorporated into the initial formulation of the model, because the fact that they describe behavior adds nothing to the justification.

We consider two counterarguments to our claim. First, one might argue that a computational model derived by embedding descriptive facts into a normative framework inherits rational properties from the normative framework. Oaksford and Chater (in press) argue that such a model will lead to 'successful thought and action in the everyday world, to the extent that it approximates the optimal solution specified by rational analysis.' (p. 32). If the model does approximate the optimal solution, then we tend to agree. We do note though that approximating an optimal solution is not necessarily the best that a constrained performer can do. In principle, a model that makes systematic errors can do better than a model that approximates optimality as long as the degree of error of the first model is small relative to the closeness to optimality of the second model. More crucially, we reject the rational inheritance argument because new facts without normative justification have the potential to lead a normative model completely astray; its recommendations for action can potentially be spectacularly wrong.

To take a prominent example in the analysis of belief, one might argue against the received wisdom that people's probability judgments are non-Bayesian (Tversky & Kahneman, 1983) by softening assumptions about what a Bayesian model should compute. One might argue that in fact judgments are Bayesian, people simply use unexpected prior probabilities in their computations (Tenenbaum & Griffiths, 2001a). We set aside whether people's judgments can actually be explained this way and focus instead on whether such an account would rationalize judgments were it able to fit the data. Clearly such an account, if valid, would reveal systematicity in people's judgments and would in that sense provide a valuable descriptive model. Were such a model descriptively valid, we might even say that people are 'coherent' in the sense that some beliefs could be derived from others (given the model).

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 484

However, such beliefs – even if coherent – would not have any rational justification, even partial justification. Prior probability distributions could be constructed that would be guaranteed to render judgments that were close to useless (for instance, a prior distribution that was the complement of true prior probabilities). The fact that people are coherent does not provide a rationalization of their beliefs if they are consistently wrong. It is easy to be coherent once you give up on accuracy.

The descriptive success of such a model does indicate that, at a coarse level, people are sensitive to the variables that the normative analysis says they should be sensitive to. For example, in the probability judgment case, the empirical success of a Bayesian model with arbitrary priors suggests that people are sensitive to priors even if the specific priors that they use have no justification. More generally, the normative framework plus descriptive facts approach does offer a method of demonstrating that people are sensitive to cues that they should be sensitive to. Clearly though the approach is overkill in this regard. An experiment that simply manipulated the relevant cue and showed that people respond accordingly would allow the same inference. There is no need to quantitatively fit model to data to find this out.

We briefly consider a second prominent example of a model of this type, the Oaksford and Chater (1994) interpretation of the Wason four-card selection task in terms of optimal data selection. On one hand, we have great sympathy for the idea that the Wason task asks participants to choose the experiments that would be informative for determining the probability of a hypothesis (a conditional rule). The Wason task is not a deductive task in that it does not ask about the deductive validity of an argument. So far, the Oaksford and Chater insight strikes us as a radically new and persuasive normative model of the task (a rational analysis of the first type). On the other hand, Oaksford and Chater go beyond this rational analysis by invoking the rarity assumption. This assumption turns out to be necessary to explain many of people's selections. The assumption of rarity makes sense in some contexts (the probability of encountering a non-black thing is indeed incomparably higher than the probability of encountering a black raven). But it doesn't hold in general; specifically, it doesn't hold in the abstract Wason four-card selection task (relative to one another, neither vowels, nor consonants, nor odd or even numbers are rare). Importing the principle into the model therefore has only a descriptive motivation, not a normative one that holds in general. The resultant model loses its rational basis once this assumption is made. Assuming rarity can cause the selection of non-optimal data.

A second counterargument to our claim that the normative framework plus descriptive facts approach does not provide a rational analysis is that, sometimes, we can only discover people's goals by examining their behavior. Once we see what they have done, we can find a reason for it; i.e., we can construct a justification after the fact. This position completely obscures the normative/descriptive distinction. Some justification can indeed always be constructed after the fact but that justification has no warrant to be called normative. Maybe people's goals are not what we expected them to be, but we cannot determine their goals and evaluate their procedure for obtaining them from the same data. People are guaranteed to come out smelling of roses that way. We might be able to determine what people's goals are and how they arrive at them via a modeling process like this, but in that case our method serves

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 485

merely as a heuristic for constructing descriptive theories, not a rational justification for behavior. Of course, there's nothing wrong with constructing descriptive theories. But calling them 'rational' just confuses the issue.

In sum, the normative framework plus descriptive facts approach to rational analysis is not a rational analysis at all in the sense of having any normative justification. The modeling framework used to perform this kind of analysis is not valid by hypothesis. The hypothesis in this case concerns an aggregate of normative and descriptive assumptions and thus cannot be shown valid by an analysis of the situation; performance itself must be considered. Therefore, what is learned is relative to independent empirical demonstration. Descriptive models can and have emerged from this enterprise, but any claim to rationality is circular.

Are causal reasoning and learning rational?

Reasoning

Rational reasoning

Normative models of causal reasoning have attempted to justify causal claims using counterfactual reasoning (Lewis, 1973), probability theory (Suppes, 1970), and a calculus of intervention (Pearl, 2000; Spirtes *et al.*, 1993; Woodward, 2003). Despite the divergence of views about the foundations of causal reasoning, we believe that philosophers and psychologists would mostly agree on what constitutes good causal reasoning for the types of cases that we will focus on.

When reasoning about a well-defined situation, people tend to be highly sensitive to the structure of the causal relations connecting events. An elegant demonstration of this can be found in Cummins *et al.* (1991) and Cummins (1995) who asked people to evaluate the strength of conditional arguments such as

> If I eat candy often, then I have cavities.
> I eat candy often.
> Therefore, I have cavities.

Eating candy is a cause of having cavities but it is neither sufficient (regular tooth brushing will prevent cavities) nor necessary (drinking soda pop can lead to tooth decay even in people who don't eat candy). So the causal relation invokes both disabling conditions and alternative causes. If people evaluate conditional arguments by reasoning causally – i.e., in a way that respects actual causal constraints that are generally known but not mentioned in the argument – then they will not find an argument such as this highly compelling despite the fact that it seems to conform to a valid logical schema, modus ponens. And people don't find it highly compelling. From this and a number of related arguments, Cummins and her colleagues conclude that human reasoning is more sensitive to causal content than to syntactic form. A corollary of this conclusion is that people are effective at reasoning with causal structures that involve a variety of disabling conditions and alternative causes. This explains why people are so good at understanding conditional statements despite the range of causal structures to which they can make reference (Bennett, 2003).

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 486

Another normative constraint on causal reasoning is the need to distinguish intervention and observation (Pearl, 2000; Spirtes *et al.*, 1993; Woodward, 2003). Observing an event licenses inference about that event's causes. For instance, observing Team A beat Team B provides evidence that Team A is stronger than Team B (and therefore, say, more likely to beat Team C). Intervening to produce the event blocks such diagnostic inferences. If I drug Team B in such a way that they are guaranteed to lose to Team A, then the loss no longer provides evidence that Team A is stronger than Team B because the loss has an overwhelmingly strong alternative explanation, namely my malevolent behavior. Strong interventions on an event cut the diagnostic link from the event to its normal causes, eliminating some of the inferences that would otherwise be possible. Understanding that intervention introduces temporary independence between an effect and its cause is particularly important in the context of decision making because choice is related to intervention (Meek & Glymour, 1994; Sloman & Hagmayer, 2006).

People are exquisitely sensitive to the logic of intervention. Sloman and Lagnado (2004) gave people simple and more complicated scenarios with well-specified causal structures and then asked them to make inferences about the scenarios with counterfactual conditionals. When the conditionals involved imagined interventions (e.g., if the effect had been prevented), people were much less likely to infer a change in the state of the cause than they did when the conditional involved an imagined observation (e.g., if the effect had been observed to not occur). People distinguished interventions from observations whether the relations were deterministic or probabilistic. The difference did not arise when the scenarios involved logical rather than causal relations. Waldmann and Hagmayer (2005) report supportive data in the context of reasoning and Hagmayer and Sloman (2007) find evidence for sensitivity to intervention in the context of choice. Even rats are sensitive to the logic of intervention (Waldmann & Blaisdell, this volume).

Another normative property of causal reasoning is screening-off (Pearl, 1988). Screening-off relates statistical dependence and independence between variables to structural relations among causes. The simplest case of screening-off arises in a causal chain. If A causes B and B causes C and if the value of B is fixed, then A and C are independent, $P(C|A,B) = P(C|B)$. In other words, the influence of A on C is mediated by B. Therefore, if the value of B is constant, A and C must be unrelated. Blok and Sloman (2006) tested people's sensitivity to this logic by giving people questions like the following:

a. The power strip isn't working, what's the probability that the computer isn't working?
b. The outlet and the power strip aren't working, what's the probability that the computer isn't working?

On the assumption that the relevant causal model for both of these cases is
Power flow in the outlet → Power flow in the power strip → Power flow in the computer

Screening-off would dictate that the answer to the two questions should be the same. Accordingly, Blok and Sloman found that the probability judgments did not differ significantly. Moreover, most people agreed with the causal model. Note however that when the task was changed slightly, when people had to choose the case with

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 487

the higher probability for the computer functioning, they chose a. more often than b. Perhaps because of the presence of an apparently irrelevant cause, knowing the outlet didn't function made the causal argument that the computer wouldn't function seem weaker. Chaigneau *et al.* (2004) also report violations of screening-off with chain structures but their stimuli came from a different domain. They described common objects by their causal structures and then asked people to name them or to infer their function.

Systematic error in causal reasoning

Several theorists have also argued that screening-off should occur for common cause structures like the following (Pearl, 1988, 2000; Spirtes *et al.*, 1993, but see Cartwright, 2002, for counter-arguments):

Power flow in the lamp → Power flow in the outlet → Power flow in the computer

In this case, the claim is that fixing the power flow in the outlet renders power flow in the lamp and the computer independent. Any statistical relation between the lamp and the computer's power flow is due to their common cause, the outlet. So if power flow in the outlet is fixed, there is no remaining source of covariation between the lamp and the computer.

Blok and Sloman (2006) tested screening-off with common cause structures using questions with the general form of a. and b. above. Violations were rampant. Knowing one effect did not occur reduced the probability of the other effect even when people knew that the common cause hadn't occurred. This happened using both rating and choice tasks. Walsh and Sloman (in press) obtained parallel results using different materials and a different judgment task. Walsh and Sloman argue that the violations are due to the nature of the explanations that people generate for whatever facts are presented. Rehder and Burnett (2005) found small violations of screening-off when they asked people to make inductions about the properties of category members.

People are known to make other systematic causal reasoning errors. The simplest illustration is that people tend to believe that a cause provides more evidence for an effect than an effect does for a cause. For example, people estimated that the conditional probability that a girl has blue eyes given that her mother has blue eyes is higher than that a mother has blue eyes, given that her daughter has blue eyes (Tversky & Kahneman, 1980; Weidenfeld *et al.*, 2005). This is rational if and only if the marginal probability of the effect is greater than that of the cause. But this is implausible in the example at hand. Blok and Sloman (2006) made a parallel observation in the context of probability judgments using questions of the type illustrated above.

Another type of error that can occur is a tendency to rely too much on a single causal model thereby neglecting alternatives. This tendency has been observed during performance of a number of tasks like probability judgment (Dougherty *et al.*, 1997) and troubleshooting (Fischhoff *et al.*, 1978). Chinn and Brewer (2001) presented people with an article reporting a theory of a scientific phenomenon. The arguments in favor of the theories were either weak or strong in terms of the amount of evidence and credibility of the sources. Next, participants were shown an article from proponents of the other theory. Articles read first proved to be more convincing regardless of the strength of the argument. The common explanation for all these phenomena is

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 488

that people make inferences from the causal model that they find most plausible at the moment of judgment rather than moderating their inferences by taking other reasonable causal possibilities into account.

The logic of causal intervention is also violated by some instances of self-deception. Quattrone and Tversky (1994) asked a group of students to hold their arms in very cold water for as long as they could. Half of the group was told that people can tolerate cold water for longer if they have a healthy type of heart, while the other half was told that the healthy heart causes decreased tolerance. The first group lasted longer than the second. This kind of behavior is not consistent with causal logic because changing tolerance for cold would not change the type of heart one has and all participants must have known this. In fact, the result was obtained even for those who denied that their behavior was affected by their knowledge of the hypothesis. Their claim that they were not affected by the hypothesis (p) along with the demonstrable fact that they were affected by it (not p) satisfies the usual definition of self-deception (Sackeim & Gur, 1978; Sahdra & Thagard, 2003; Talbott, 1995). Bodner and Prelec (2002) ascribe cases of self-deception to self-signaling. People violate the logic of causation in order to signal information about themselves to others and also to themselves. Quattrone and Tversky's (1984) participants aimed to signal to themselves that they possessed a good heart.

Other violations can result in tangible loss. Many people co-operate in a one-shot prisoner's dilemma game (e.g., Morris *et al.*, 1998; Shafir & Tversky, 1992). Co-operation is inconsistent with a desire to maximize the utility of causal consequences (Nozick, 1995) because whatever the opponent does, a player does better by not co-operating. In a related vein, many people choose one box when put in a Newcomb's Paradox situation even when choosing two boxes is the dominant option (i.e., it is better regardless of the state of the game). Whatever people are doing in these situations, they are not correctly determining the joint effects of their actions and the environment and choosing so as to maximize their benefits.

There are many plausible explanations why people violate causal logic in these cases. Here we focus on their justification, not their explanation. Talbott (1995) argues that even a coherent, rational agent will sometimes benefit by self-deception. Rational agents have an interest not only in obtaining favorable outcomes, but also in believing certain facts about the world whether they are true or not. For example, we all have an interest in believing that our parents love us. As long as we're not so duped that we fail to foresee some horrible event, like abuse or abandonment, we benefit by believing that we are important enough to deserve love whether or not it is true. Similarly, the belief that there is hope in our future is of value whether true or not. Talbott proves that under a set of reasonable assumptions about belief and its utility and about gathering evidence for belief, a Bayesian utility maximizer should act to maintain desirable beliefs rather than to remain unbiased, as long as the desirable beliefs do not have sufficiently bad consequences.

Mild to moderate self-deception has psychologically positive consequences (see Taylor & Brown, 1986, 1994). People have inflated views about their traits (Brown, 1986) and abilities (Campbell, 1986; Dunning *et al.*, 2004, provide a review). Those positive self-illusions elevate people's mood, their well-being, and their self-esteem

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 489

(Taylor & Brown, 1988, 1994). Moreover, self-serving signaling of capabilities strengthens people's belief in personal control (Bandura, 1977). Mild to moderate forms of self-deception can increase mental health. This helps to explain the voluminous evidence for motivated reasoning (e.g., Blanton & Gerrard, 1997; Kunda, 1990) and memory (Erdelyi, 1996). This kind of reasoning neglects causal beliefs for the sake of a desired conclusion having positive emotional consequences.

Implications for the rationality of causal reasoning

In summary, causal reasoning proves to be well described by rational principles in some respects. People also seem to understand the difference between observation and intervention. People are highly sensitive to causal structure, adjusting their inferences appropriately in the face of enabling and disabling conditions and appropriately screening-off variables when the mediating variable is fixed in a chain structure. But people also make systematic errors. They do not always screen-off when they should such as when a common cause is fixed. They are more willing to make causal inferences than diagnostic ones along the same causal path and people tend to rely too much on a single causal model. Finally, people can act in ways that violate their own causal beliefs such as in cases of self-deception, co-operation in two-player games, and certain self-serving biases in reasoning and decision making.

Our conclusion is that people indeed have the tools for effective causal reasoning, but clearly there is room for improvement. We have to remind ourselves to consider alternative possibilities and to consider with deliberation whether evidence really supports our conclusion and to what extent.

Self-deception and other acts of self-signaling are clearly errors of causal reasoning and yet they have a justification stemming from goals broader than the narrow goal of getting a particular answer correct. Indeed, we have seen that they can be justified in a Bayesian framework by considering their relative utility. But whatever process supports these acts also supports acts that have no justification. If one's goal is to maximize income, then co-operating on a single-shot Prisoner's Dilemma makes no sense nor does choosing one box in a Newcomb's Paradox (see Nozick, 1995, for explanation) once backwards-causality is ruled out (see Shafir & Tversky, 1992, for such a case). Not all acts of signaling can be justified even if some have psychological benefits.

Learning

Evidence for rationality

The debate about the rationality of inference in the context of causal learning has focused on how people learn causal relations based on contingency information. The typical experiment presents a participant with frequency data about the covariation of a putative cause (e.g., a drug) and an effect (e.g., curing a disease) either in trial-by-trial or tabular form. These data can be viewed in terms of a two-by-two table, based on the presence or absence of the cause and effect. This literature has shown that people tend to differentially weight the cells when making inferences. They pay most attention to the data in which both the cause and effect are present and the least attention

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 490

to the data in which both are absent. This has been taken as evidence for the non-rationality of human contingency judgments on the assumption that all cells should have equal weight in determining whether or not there is a causal relation.

Anderson (1990) however has argued that a normative causal inference from contingency data should differentially weight the cells and should do so in precisely the way that people do. He proposes a Bayesian model of contingency learning in which the likelihood of the data under a target causal model is compared to the likelihood of the data under an alternative model. The target causal model posits some probability of the effect in the presence of the cause and some probability of the effect in the absence of the cause. The alternative model posits that the effect occurs with some base probability that is independent of the proposed cause. According to the Bayesian learning model, the amount of evidence that each of the cells provides for one or the other causal model is a function of these three probabilities and reflects how well the data in the cell distinguishes the two models. It turns out that the cause present/effect present cell is the most diagnostic as long as the probability of the effect in the absence of the cause in the target model is close to the base probability in the alternative model. To see why, consider an example: Suppose I want to determine whether eating a particular salad dressing causes me to have an upset stomach. I might assume that if the salad dressing is the culprit, most of the time that I eat it I will feel unwell. Presumably I am unlikely to have an upset stomach whether the salad dressing does make me ill but I don't eat it or if the salad dressing is not the cause of my illness. Under these conditions, eating the salad dressing and getting sick provides strong evidence for the causal model in which the salad dressing is a cause of illness because under the alternative model, such an outcome is unlikely. Conversely, not eating the salad dressing and feeling fine provides little evidence to discriminate the models. Under both models such an outcome is likely. Thus the differential cell weighting can be interpreted as evidence that people's causal inferences from contingency data are informed by a rational inference rule that gives greatest weight to data that are informative given the task. The key assumption is that the task is not one of determining the raw probability of a causal relation, but of comparing two causal models.

The developmental literature also reveals some support for a causal learning mechanism that is guided by normative principles. Between the ages of three and four, children begin to make causal inferences that are consistent with a Bayesian prescription. For instance, 4-year-olds are sensitive to backwards blocking and explaining away relations. Sobel *et al.* (2004) performed a series of 'blicket detector' experiments with three and four year olds in which children observed certain objects or combinations of objects activating a machine (the 'blicket detector') and were asked to infer whether a particular object would do so (i.e., whether it was a 'blicket'). In the explaining-away condition, 2 objects, A and B, activated the machine together. In a subsequent trial, A failed to activate the machine alone. Participants were asked to infer the identity of B. As A alone did not elicit the effect, participants should have inferred a causal relation between B and the detector. In the backwards blocking condition, A and B again activated the detector together but this time A also activated the machine alone. Because a causal relation between A and the detector provided an explanation for the effect in the first trial, participants should have been less likely to infer a causal relation

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 491

between B and the detector than in the explaining away condition. Indeed, 4-year-olds were likely to assert a causal relation between B and the detector in the explaining-away condition, but not in the backwards-blocking condition. In another experiment it was shown that four year olds' inferences in the backwards-blocking condition can be modulated by varying the base-rate of objects that are 'blickets', again consistent with rational norms. These findings demonstrate some rationality in causal learning in young children.

Evidence against rationality

Not all aspects of causal learning have a rational justification. Even within the contingency learning paradigm there is some counter-evidence in the way that participants respond to sample size. As the number of observations increases, differences in the probability of an effect in the presence and absence of a putative cause increase support for a causal relation because these differences are proportionately less likely to be due to chance. However, Anderson and Sheu (1997) found that participants were not sensitive to sample size when inferring causal relations from contingency data.

Moreover, when more complex causal learning tasks have been investigated, further problems arise for a rational account. In both the contingency learning and blicket detector paradigms the experimental task is relatively simple. In contingency learning, the structure of causal relations is evident. There is one putative cause and one effect and the task is to determine whether there is a causal relation from the former to the latter. Likewise, experiments in the developmental literature typically define the putative causes (e.g., actions on objects) and effects (e.g., responses by a detector). Causal relations are often deterministic and experiments consist of very few trials. Under these conditions people perform quite well. However, real world causal learning often has many potential variables and no *a priori* structure.

Some recent work has examined more complex causal learning paradigms. Lagnado and Sloman (2004) had participants observe data from a three-variable causal model and asked them to infer the causal structure responsible for generating the data. In one of the cover stories the causal structure represented a chemical process for production of a perfume and the causal variables were the presence of an acid, the presence of an ester and the ultimate creation of the perfume. On a given trial a participant might observe that the acid and ester were present and the perfume was created. On a subsequent trial he or she might observe that the acid was present, the ester was absent and the perfume was not created, and so on. Steyvers *et al.* (2003) used a similar paradigm in which the causal structure was represented by the mind-reading capabilities of aliens. In contrast to contingency learning, experiments of this type necessitate consideration of many more possible causal relations and frequency information cannot be summarized in a two-by-two table. The results of these experiments are striking in the degree of sub-optimality they reveal. Given just observation data, people rarely infer the model actually generating the data. When allowed to make interventions, performance improves, but still surprisingly few people are able to reconstruct the model that produced the data.

One of the reasons for people's difficulty may be that these experiments provide few cues to causal structure beyond covariation and there is little evidence that people are

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 492

capable of tracking the statistics of the data presented during these tasks (Lagnado *et al.*, in press). The contingency learning literature shows that given access to the necessary data, some participants can make inferences in line with normative models, but in causal structure learning experiments limitations in tracking covariation online may lead participants to rely on other sources of information to make inferences.

One important cue to causal structure is temporal contiguity. Events that happen close in time are often causally related, and the direction of the relation is informed by their temporal order (effects never precede their causes). Lagnado and Sloman (2006) have shown that people easily learn causal structure from temporal cues and that temporal cues often trump covariation as a source of information for causal learning. In one experiment, participants observed virus propagation among computers in a network. On a given trial, each computer was either infected or clean, and after observing many trials the participant was asked to identify the pathway that viruses were being transmitted along. Covariation data was consistent with a particular causal structure but the temporal order was manipulated to favor a different structure. (i.e., participants were told that a virus could be transmitted from one computer to another, but show up in the second computer before the first). The only way to recover the true structure was to ignore the temporal information and use covariation. Even though participants were told that the temporal information was uninformative they still tended to make inferences consistent with it and inconsistent with the covariation data. White (2006) also found that participants tend to rely on temporal information rather than covariation when making causal inferences. Hagmayer and Waldmann (2002) and Buehner and May (2002) have shown that people can adjust their inferences to the expected time scale of events. This provides further evidence that people are highly sensitive to temporal information.

Another important source of information is background knowledge that constrains the set of plausible causal structures. For example, if my washing machine started making horrid noises I might reason that a broken part was responsible, and infer that replacing the part might fix the problem. I would be unlikely to make the same inference if my child were making horrid noises. One of the main findings of research in the scientific reasoning paradigm is that children and adults' theoretical beliefs about causal structure bias the way that new information is interpreted (Kuhn & Dean, 2004). In scientific reasoning experiments a participant is given covariation data for several putative causes and an effect and must infer the actual causes. Participants tend to rely on their beliefs about the mechanisms responsible for the causal relation rather than the covariation information when making inferences. For example, participants asked to identify the causes of the success of a children's show sometimes choose humor even if such an inference is not supported by the data.

A key difference between cues such as temporal order and background knowledge on one hand and covariation on the other is that the former are available on individual trials whereas covariation is based on integrating over multiple trials. Tracking covariation is often difficult and therefore other sources of information are used. Much of causal learning may use simplifying heuristics that are sensitive to trial-to-trial cues (Fernbach, 2006; Lagnado *et al.*, 2007). Under many conditions, these heuristics approximate normative inferences but they can also result in systematic errors.

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 493

Fernbach (2006) offers evidence that trial-to-trial cues are used to learn causal structure. Participants observed interventions on a three-variable causal system composed of binary slider bars and had to infer the causal model generating the data. For each causal model, participants observed five interventions, and prior to each one the intervened-on slider bar was identified. Thus the intervention provided a trial-to-trial cue. The participant knew that the intervened-on slider bar moved first and that if any other slider bar moved it was a direct or indirect effect. The majority of participants' responses can be explained by a simple heuristic that asserts a direct causal relation between the intervened-on variable and any other variables that were activated during that intervention. One characteristic result was that when the generative model was a chain (e.g., slider A causes slider B and slider B cause slider C) few participants ever inferred chains. Instead they tended to infer confound models, three-link models that include links from the root variable to both of the other variables and a link from the second variable in the chain to the third (e.g., A causes B, A causes C, and B causes C). Use of a heuristic based on the interventional cue offers a simple account of this result. In the case of an A causes B causes C chain, a participant will rarely get evidence for A to B and B to C links without simultaneous evidence for an A to C link because A and B will rarely both be active in the absence of C. Participants were thus unaware that the link from the second variable to the third was unnecessary to explain the data and systematically inferred extraneous links.

Several counter-normative phenomena of human cognition may be due to the sub-optimality of simplifying heuristics used for causal learning. For instance, both animals (Skinner, 1947) and people (Ono, 1987) have been shown to infer a causal relation between unrelated events based on few or even a single observation of temporally contiguous events. This type of learning may be the source of superstitious behavior. Another example mentioned above is that people tend to be biased in interpreting evidence to accord with pre-existing beliefs about causal mechanism. Such biases may stem from the relative ease of accessing and using background knowledge as opposed to covariation data.

Implications for the rationality of causal learning

In summary, people learn causal relations in a way that is consistent with some rational principles from the age of 4 or younger. In contingency learning, people may differentially weight types of evidence in a way that allows the diagnosis of competing models. Likewise, 4-year-olds are sensitive to rational principles such as explaining away and to base rates. But even in these relatively simple tasks there is evidence of systematic bias, such as sample size neglect. When the task is more complex, people are incapable of tracking covariation and using it to make effective statistical inferences about causal relations. Thus it is unlikely that the dominant mechanism for real world causal learning is anything like a system that tracks the covariation of events and performs a statistical inference over the data. Rational models of the type proposed by Anderson (1990), Tenenbaum and Griffiths (2001b), and Gopnik *et al.* (2004) are probably not accurate as comprehensive accounts of causal learning.

Rather, causal learning is often based on heuristics sensitive to cues other than covariation such as temporal information, background knowledge, and interventions.

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 494

This is similar in spirit to Anderson and Sheu's (1995) claim that 'subjects look for quantities that are easy to compute and which are causally relevant and make their judgments with respect to these' (p. 39) and Kuhn and Dean's (2004) claim that the strategies people use to solve causal inference problems vary within and across individuals depending on the nature of the task. Though these heuristics are generally effective they do lead to systematic biases.

Conclusion

A number of important models of reasoning and learning have come out of the tradition that associates itself with rational analysis (e.g., Oaksford & Chater, 1994; Steyvers *et al.*, 2003). The models are computational in the sense that they hypothesize a functional relation that people are trying to accomplish. But the models are descriptive, not normative. The models include assumptions chosen to fit data, not only to satisfy task constraints or resource limitations, and in that sense have no claim to rationality. Rational frameworks offer a rich menu of mathematical tools that should be made use of by theorists, but if the model includes assumptions chosen to fit the data, the model no longer has a rational basis.

Some theorists argue that a framework that supports coherent reasoning has a kind of rationality even if it includes false beliefs. But reasoning logically about false statements does not lead to truth. Analogously, reasoning coherently about false beliefs is not a form of rationality. It might be rational to run away from a tiger, but running away from a bunny rabbit is not rational even if you believe the rabbit is a tiger.

We reiterate that our goal is not to damn the descriptive value of the kind of models that have been and will be developed in the rational analysis tradition. In fact, we suspect that models of reasoning and learning will always require some notion of maximizing performance to account for human flexibility and adaptiveness. This is obvious when modeling thinking in the real world. People learn (eventually) to fix their bicycles effectively. But this kind of display of everyday rationality does not entail that normative models serve as a prototype for descriptive models. It merely tells us that descriptive models require a quality control or error-reduction mechanism. Normative models may inspire hypotheses about what those mechanisms might look like, but the normative and descriptive models remain distinct entities and we should identify them that way in order to maintain coherence and clarity in our theorizing.

Our reviews of causal reasoning and learning have revealed some of the reasons that rational models have had limited descriptive power; others are reviewed by Danks (this volume). Although causal reasoning can be highly effective and demonstrates exquisite sensitivity to causal considerations, it shows some definite deficiencies. In order to solve a huge variety of complex and subtle problems, it has developed strategies that lead to systematic bias. Similarly, causal learning can pick up on a variety of trenchant cues to figure out the causal structures that relate events. Yet it fails to pick up on some of the most valuable information, like covariational data, and it often guesses using a smorgasbord of strategies. Treating people as rational may not be irrational but it is a mistake; it fails to identify some of the aspects of cognition that are uniquely human.

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 495

Acknowledgements

This work was supported by NSF Award 0518147. We thank York Hagmayer for useful discussion.

References

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

Anderson, J. R., & Sheu, C.-F. (1995). Causal inference as perceptual judgments. *Memory & Cognition*, **23**, 510–524.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford University Press. Oxford, UK.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning*, **8**, 269–295.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, **53**(3), 411–453.

Chaigneau, S.E., Barsalou, L.W., & Sloman, S. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, **133**, 601–625.

Cherniak, C. (1986). *Minimal rationality*. Cambridge: MIT Press.

Cummins, D.D. (1995). Naive theories and causal deduction. *Memory and Cognition*, **23**, 646–658.

Cummins, D.D., Lubart, T., Alksnis, O. and Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition*, **19**, 274–282.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, **70**, 135–148.

Fernbach, P. M. (2006), Heuristic causal learning, *First Year Project, Brown University*.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representations. *Journal of Experimental Psychology: Human Perception and Performance*, **4**, 330–344.

Gopnik, A., Glymour, C.,Sobel, D. M.,Schultz, L. E., Kushir, T., & Danks, D. (2004). A Theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, **111**, 3–132.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (in press). Bayesian models of cognition. In R. Sun (ed.), *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, **30**, 1128–1137.

Harman, G. (1995). Rationality. In E. E. Smith & D. N. Osherson (Eds.), *Thinking (an invitation to cognitive science* (Vol. 3). Cambridge: MIT Press.

Kuhn, D., & Dean, Jr., D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development*, **5**(2), 261–288.

Lagnado, D. A., Waldmann, M. R., Hagmayer Y., & Sloman, S. A. (in press). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford: Oxford University Press.

Lagnado, D., & Sloman, S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **30**, 856–876.

Lagnado, D., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **32**, 451–460.

21-Charter&Oaksford-Chap21 11/5/07 11:04 AM Page 496

Lewis, D. (1973). *Counterfactuals*. Cambridge: Harvard University Press.

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

Meek, C., & Glymour, C. (1994). Conditioning and intervening. *The British Journal for the Philosophy of Science* **45**, 1001–1021.

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

Oaksford, M., & N. Chater (in press). Bayesian rationality: The probabilistic approach to human reasoning. Oxford: Oxford University Press.

Ono, K. (1987). Superstitious behavior in humans. *Journal of the Experimental Analysis of Behavior*, **47**, 261–271.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, **50**, 264–314.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, **69**, 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, **63**, 129–138.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Non-consequential reasoning and choice. *Cognitive Psychology*, **24**, 449–474.

Skinner, B. F. (1947). 'Superstition' in the pigeon. *Journal of Experimental Psychology*, **38**, 168–172.

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

Steyvers, M., Tenenbaum, J., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, **27**, 453–489.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland Publishing Company.

Tenenbaum J. B., & Griffiths T. L. (2001a). The rational basis of representativeness. 23rd *Annual Conference of the Cognitive Science Society*, 1036–1041.

Tenenbaum, J., & Griffiths, T. L. (2001b). Structure learning in human causal induction. *Advances in Neural Information Processing Systems* (Vol. 13). Cambridge, MA.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgment under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale NJ: Erlbaum.

Walsh, C., & Sloman, S. A. (in press). Updating beliefs with causal models: Violations of screening off. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and Mind: A Festschrift for Gordon H. Bower*. New Jersey: Lawrence Erlbaum Associates.

Weidenfeld, A, Oberauer, K, & Hornig, R. (2005). Causal and noncausal conditionals: an integrated model of interpretation and reasoning. *Quarterly Journal of Experimental Psychology A*, **58**(8), 1479–1513.

White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, **18** (3), 454–480.