Information Knowledge Systems Management 10 (2011) 1–15 1
DOI 10.3233/IKS-2012-0187
IOS Press

Chapter 5

Human representation and reasoning about complex causal systems

Steven A. Sloman* and Philip M. Fernbach
Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI, USA

1. Introduction

Sociotechnical causal systems tend to be complex and dynamic. They involve multiple elements, elements that are often themselves sophisticated, that interact in ways that can be chaotic and that can lead to sudden, catastrophic changes. Moreover, sociotechnical systems often must respond to stochastic inputs. Such systems are inherently hard to understand, control, and predict. We focus on how people think about such systems. It is critical to understand how people think about complex sociotechnical systems for two reasons: First, we must intervene on systems when they break down or when we want to improve or nudge them. The intervention we choose will depend on how we understand the operation and trajectory of the system. If our understanding is biased in any way, then our interventions might be flawed. By knowing the nature of human bias, we might be able to correct those flaws. Second, humans are players in sociotechnical systems, so predicting the behavior of the system requires understanding the behavior of the people within the system. But people are not (usually) passive nodes in a larger network. How people think about the system they are working in influences how they will act and react, and hence their contribution.

Our objective in this chapter is to characterize what is known about how people represent, reason about, and predict the behavior of complex systems. The other chapters in this collection document how big causal systems can be, what kinds of problems they face, and how much trouble people have dealing with them. We will not review that mass of data here. Rather, we will focus on the dimension of human understanding: where people go wrong – the cognitive foibles, tricks and shortcuts that determine how we understand complex systems – and on what we do well.

Our aim is to describe what human cognition brings to the table in the understanding of complex sociotechnical systems. Of course, cognition does not take place in a vacuum. Cognition is to a large extent a social enterprise, cognitive acts depending in general on a "community of knowledge" in which

*Corresponding author: Steven A. Sloman, Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI, USA. E-mail: steven_sloman@brown.edu.

knowledge is distributed across the members of a community. We also subscribe to the belief that human reasoning is "embodied" not just in an individual's own body but in physical systems that can include artifacts and other objects (cf. Barsalou, 1999). For instance, Hutchins (1995) has described how it is the cockpit and not the pilot alone that flies an airplane; all of the pilot's thought processes depend on the state of controls and displays in the cockpit and vice versa. This kind of embodiment is central to good human factors design. Our focus in this chapter however is to attempt to describe the contribution of the human mind to the study of sociotechnical systems. How the mind interacts with the relevant system will depend on the particulars of any actual case.

We will argue that people reason about complex systems by simplifying in three ways: First, people resort to a variety of heuristics that are selective in what information they consider. These heuristics often yield satisfactory results though they can lead to systematic error. Second, when people do try to take more information into account, they often use a model that has a simple linear form that ignores most of the interactions and sources of unpredictability in the system. Finally, when people go beyond heuristics and simple linear combinations, they tend to build a mental model that reflects the causal structure of the system by representing the mechanisms that lead from causes to effects. We refer to such representations as causal models. Although people excel at representing how individual mechanisms work and how they are linked to each other, they tend to neglect cycles of causation, often fail to reason quantitatively, and sometimes ignore relevant variables wholesale. The nature of causal models is the most poorly understood of the three forms of simplification and will receive most of our attention in this chapter.

2. Simplifying heuristics

Often people do not try to understand a complex system but rather take an educated guess based on a rule of thumb that can be quite effective, though not precisely accurate. Many heuristics have been identified in the judgment and decision-making and social cognition literatures (Gilovich & Griffin, 2002). Here is a list of the most important ones that are used frequently and are supported by substantial evidence:

- *Representativeness*: The probability that object A belongs to class B or originates from process B is evaluated by the degree to which A resembles B (Tversky & Kahneman, 1983). To illustrate, the average citizen is more likely to expect an economic catastrophe after a banking crisis than after a real estate crash because banks deal with money and therefore seem more like players in the economy than homes do.
- *Availability*: People assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind (Tversky & Kahneman, 1973). Everybody worries about a terrorist incident just after a terrorist incident and about an earthquake after an earthquake.
- *Recognition/Fluency*: Familiarity breeds probability. Events are judged more likely if they are recognizable or easy to process and therefore seem familiar. For instance, machinery or tactics that are familiar seem more likely to work even in environments in which they have not been tested. A boundary condition on this effect is when events are unfamiliar but a little thought will bring to mind a good reason for the event to be probable. In that case, disfluency can cause people to think harder about the event so that they find that good reason and judge the event more probable (Alter & Oppenheimer, 2009). Disfluency can cause people to deliberate more.

- *Anchoring & adjustment*: People often make estimates by starting at an initial value and adjusting to yield a final answer. Adjustments are typically insufficient (Tversky & Kahneman, 1973). For instance, forecasts of the length of a military campaign often anchor on a previous similar campaign. While forecasters will try to adjust for the unique properties of the current campaign, those adjustments tend to be insufficient.
- *Simulation heuristic*: Events that can be mentally "simulated" with ease are judged more likely (Kahneman & Tversky, 1982; Wells & Gavanski, 1989). The claim here is that, regardless of the quality of a mental simulation (discussed below), the mere fact of constructing a simulation of an event appears to raise its probability. For instance, many people feel less safe in airplanes than in trains because it is easier to imagine a heavy metal object falling out of the sky than the derailment of a train.
- *Proximity heuristic*: People use judgments of closeness (distance) to estimate risks and probabilities (Teigen, 2005). Fear of terrorism is often proportional to proximity to the location of the latest attack with a resultant neglect of the other factors determining a terrorist's choice of target.
- *Take the best*: When multiple dimensions are relevant to a judgment, the dimensions are tried one at a time according to their cue validity, and a decision is made that favors the first dimension that discriminates the alternatives (Gigerenzer & Goldstein, 1996). Stocks are often purchased based only on past performance without regard to changing financial or management conditions of the firm.

Each of these heuristics is effective in the sense of capturing an important property of the event, a property highly correlated with the probability of the event. Thus these heuristics often lead to very accurate explanations, forecasts, and predictions. However, each can lead to systematic errors when the information captured by the property fails to be predictive. For instance, California has warm weather on average but predicting the weather in Berkeley, California on this basis will fail miserably because Berkeley lies within a microclimate that makes it much colder on average than one might expect. A second illustration: Unsecured baggage in an airplane is a much greater danger to one's safety than the plane falling out of the sky, though it is rarely of concern to passengers (see Kahneman, 2011, for many examples).

The heuristics and biases framework has focused on simple judgments and predictions of single values but some work has explored judgments concerning more complex data, namely sequences (Gilovich, Vallone, & Tversky, 1985). These studies suggest that people impose causal structure in their interpretation of even random sequences (more on this below).

3. Simple linear combinations

The point of a heuristic is to capture some invariant in the behavior of a complex system that provides predictive power while ignoring most available information like the relevant system variables themselves or their relations to one another. One important question about the relations among variables that heuristics do not consider explicitly is how they trade-off, how one variable compensates for another. For instance, judgments regarding investment strategies, marketing programs, energy resources, military campaigns, and purchases require trading off costs with potential benefits. People are often very concerned about trade-offs. In such cases, rather than using strategies that focus on a single dimension, people will frequently rely on strategies that take trade-offs into account by offering a means by which a high value on one dimension can compensate for a low value on a different dimension.

In complex systems, variables tend to trade-off in complex, non-linear ways. Nevertheless, people tend to use linear combinations to represent trade-offs (Anderson, 1981; Brunswick, 1955; Hogarth & Einhorn, 1992). Linear combinations of predictor variables often correlate more highly with human judgments in situations involving complexity than models that represent the actual structure of the situation (Dawes, Faust, & Meehl, 1989). In fact, models that give each variable equal weight tend to do just as well as linear models that involve some kind of regression to optimize the weight assigned to each variable in the linear combination (Dawes, 1979). Generally, human judgment can be modeled by merely normalizing the variables; they need not be differentially weighted.

The most surprising fact about these kinds of linear models (even "improper" linear models giving equal weight to each variable) is how effective they are under normal circumstances. When the environment cannot be modeled with a reasonable degree of certainty, they do just as well, or better, than more sophisticated representations. This is especially true in clinical domains (a domain without a good model; e.g. most mental health and education domains, predicting recidivism rates, or academic or job performance). In such domains, linear models almost always do better than expert judgment (Dawes, Faust, & Meehl, 1989). That is, the predictions of an expert who is given all available evidence will be further off the mark than a linear model based on the variables the expert him or herself considers important. Of course, this suggests that experts do not use linear models as a matter of course, or their predictions would be at least as accurate as linear models. Often, experts rely on simplifying heuristics or on causal models, to which we now turn.

4. Causal models

Experts generally are able to recognize rich patterns within their domain of expertise. These patterns often reflect causal beliefs (see Sloman, 2005) and are often communicated as causal narratives. Consider a decision-maker determining, say, a company's management strategy. The decision-maker needs to represent the structure of the company, how it produces products, and its market environment. This requires representing a complex system of causal relations because it is causal laws that govern complex systems and carry them into new states. There are causes of the behavior of actors, causal mechanisms that produce products, and causes and effects of income, sales, growth, changing markets, and competition, etc. A model of those causal mechanisms thus allows for the prediction of future states by offering a representation that affords mental simulations, for instance simulations of different marketing strategies to evaluate their effectiveness.

Beyond representing the problem to allow choosing the best course of action, the decision-maker needs to justify the decision to all stakeholders, including owners, management, and labor. The decision-maker also needs to justify the decision to him or herself. A story provides a means to do that; it is a medium for representing a complex causal structure whose states change dynamically in a way that people can understand. So the question of how people represent and communicate about complex systems is not that different from the question of what makes a good story (see Pennington & Hastie, 1993).

4.1. Causal Bayes nets

The most common way to model a causal system in psychology is as a graphical representation of a probability distribution called a causal Bayesian network (causal Bayes net; Pearl, 2000; Spirtes, Glymour, Scheines, 1993). A Causal Bayes net has two components: (1) qualitative structure in the form of an acyclic graph composed of nodes and links and (2) quantitative parameters in the form

of probabilities and conditional probabilities. Nodes represent variables and directed links (arrows) represent probabilistic dependencies. Two variables are statistically independent if their nodes have no directed pathway connecting them and they do not share any parents (a parent is a node upstream on a directed path). In a causal Bayes net, the links do not merely represent statistical dependence but also represent causal power in the sense that they support intervention (Sloman, 2005, offers a simple non-mathematical introduction and Woodward, 2003, offers a philosophical analysis). A variety of theorems and algorithms have been developed for Causal Bayes nets that allow for correct probabilistic inference under very general conditions (Pearl, 2000).

The obvious limitation of Causal Bayes nets for representing complex systems is that complex systems need not be acyclic. In theory, this is not problematic because a cyclic network can be approximated to any degree of precision by unfolding an acyclic, hierachical network over multiple time steps (cf. Pearl, 1988). In this sense, Causal Bayes nets offer a general framework for the representation of causal structure. Nevertheless, they are not necessarily useful for representing complex systems as they fail to make all properties of such systems transparent. For instance, a causal model does not necessarily make explicit the stable states or bifurcation points of a system.

However, this is not a problem for our purposes because the goal of a psychological model is a faithful representation of human understanding, not the best representation of the system itself. If people do not explicitly represent the stable states of a dynamic system then the model should also not represent them. And, in the case of explicit reasoning, there are definite limitations on the complexity of what people can represent (e.g., Gentner & Stevens, 1983; Hegarty, 2004; Rozenblit & Keil, 2002). Although people are able to master complex skills (like natural language and guitar playing) over extended periods of time, there is no evidence that we are capable of explicitly representing and reasoning about the complex dynamics of a system over a shorter period of time in the absence of a formal tool to guide us. Our sensory systems and some intuitive reasoning are more responsive to dynamic structure (Elman et al., 1996). When it comes to deliberate reasoning and decision-making, processes that depend on attention and working memory, people are better at representing static than dynamic structure. This is one reason language can mislead or even deceive, because it elicits static structures to represent dynamic systems. For instance, a term like "causal model" is misleading in that it suggests a static representational structure when in fact mental representations change constantly as the environment changes, as knowledge is updated, and as goals change.

4.2. How people reason about simple causal structures

People think locally about causal systems and they do so, for the most part, very effectively. They are able to reason about the mechanisms that lead from causes, disablers, and enablers to effects. The evidence we will review in this section suggests that they do so by mentally simulating processes as they occur over space and time to produce effects.

Variables can be related through various causal structures. For instance, A may be a cause of B (A is 'predictive' of B), an effect of B (A is 'diagnostic' of B), or A and B might have a common cause. Different causal structures license different inferences. For instance, if A is the sole cause of B, then B may not be guaranteed to occur if A does (because B might be disabled by a third variable C), but A is guaranteed to have occurred if B does. People are very good at making these kinds of distinctions. They distinguish whether an inference is predictive, diagnostic, or from a common cause (Bes, Sloman, Lucas, & Raufaste, 2010; Fernbach, Darlow & Sloman, 2011a; Hagmayer & Sloman, 2009; Waldmann & Holyoak, 1992).

We can also see this kind of qualitative sensitivity to structural implications by examining whether human reasoning follows the dictates of causal logic. Perhaps the most critical feature of causal models is that they support inferences about intervention. In that sense, causal reasoning is an extension of action. The value of action is that it allows us to set a variable to whatever value we want within physical constraints. By doing so, we determine that variable's value and nothing else does. In other words, the variable is rendered independent of its normal causes. This simplifies the causal structure because the links to the variable from its causes can be treated as absent (Pearl, 2000; Spirtes, Glymour & Scheines, 1993). In the philosophical and computer science literatures, this is known as 'edge-breaking' or 'graph surgery.' This simplification does not occur when variables are merely being observed and so situations can be set up in which the inference from the state of one variable to another differs depending on whether the first variable is intervened on or observed. Parallel preparations can be done without physical intervention, by examining mental interventions. People can be asked to imagine counterfactual situations by asking them to intervene on their causal models. People draw appropriate inferences in such situations providing evidence that they understand the logic of intervention (Hagmayer & Sloman, 2009; Kaufmann, 2009; Sloman & Lagnado, 2005). So do rats (Blaisdell, Sawa, Lesing & Waldmann, 2006). Rips (2010) discusses an alternative view.

Another property of causal inference is that if an intermediary variable's value is known, then variables on either end of the intermediary are rendered independent. This is referred to as "screening-off". This property supports inference in both a common cause situation (more than one effect of the same cause) and in chains of causation. For instance, if one domino falls but the domino it hits is supported and therefore does not fall, then dominoes downstream from the supported domino are rendered independent of the falling domino (this holds even if the dominoes are spaced unevenly so that each reaction is only probabilistic). Human reasoning is sensitive to this pattern and often appeals to it, although sometimes violates it especially in the common cause case (Park, 2011). When people learn that one effect of a cause has not occurred, they tend to think other effects are less likely even if they know the cause has occurred. One reason for this is that people tend to add variables that they are not told about into their representations. For instance, if they know a cause occurred but one of its effects did not, then they will often introduce a disabling condition for the effect. That disabling condition might also disable other effects (Rehder & Burnett, 2005; Walsh & Sloman, 2008).

A third property of causal inference is the 'discounting' or 'explaining away' of one putative cause of an outcome when another sufficient cause is known to be present (Kelley, 1972; Morris & Larrick, 1995). If an effect has two independent causes that are each sufficient for the effect, then knowing that the effect occurred increases the probability of the causes. But if one also knows that one of the causes has occurred, then the probability of the other cause decreases. The known cause explains the data (the effect) and thus explains away the other cause. Although people discount, they do not always discount an appropriate amount. For instance, in accounting for others' behavior, we tend to attribute too much to personality characteristics and not enough to the environment in which the behavior occurred. This can lead to insufficient discounting of personality but too much discounting of environmental influences (Jones & Nisbett, 1972). There are also cases where people treat independent causes as if they are mutually exclusive and discount too much (Kun, Murray & Sredl, 1980).

In sum, people are able to construct small, local causal structures online and reason about them qualitatively. However, their inferences are not always quantitatively accurate; they do not in general correspond in detail to the dictates of probability theory.

4.3. Mental simulation in predictive inference

People tend to neglect alternative causes in predictive inference (Fernbach, Darlow, & Sloman, 2010, 2011b, Fernbach & Rehder, 2011). For instance, when told that a patient is suffering from a disease and asked what the probability is that the patient has a particular symptom, doctors report the probability of the symptom given the disease assuming the absence of all other possible causes of the symptom. They do not do this when reasoning diagnostically, from symptom to disease. When told that a patient has a particular symptom and asked what the probability of a disease is given the symptom, doctors tend to respond appropriately, by balancing the target disease against other potential causes.

The fact that people neglect alternative causes when making predictions has led to a counterintuitive result: Weak supporting evidence for an outcome reduces the judged probability of that outcome (Fernbach, Darlow, & Sloman, 2011a). For instance, we told a group of people in early 2010 about an endorsement in a local paper of a single Republican candidate. We then asked this group, as well as a group not told about the endorsement, whether they wanted to gamble on the Republicans winning the 2010 mid-term elections. Even though everyone agreed that the endorsement made the Republicans (slightly) more likely to win, those told about the endorsement were less likely to take the gamble than those not told about it. Even though the people themselves thought the evidence made the event more likely, they were less likely to bet on it. We believe this occurred because those given the weak cause neglected other causes whereas those not given the weak cause thought of stronger ones themselves.

These results suggest a particular model of how people make predictions or, more generally, how we reason in a forward causal direction: We imagine a local mechanism specifying how the particular cause would lead to the effect rather than thinking about the entire causal system and run a mental simulation of that specific mechanism rather than considering the whole system. This idea is consistent with how people think about what it means to be a cause, as we will see below.

4.4. Learning causal structure

People are effective at learning simple causal structures if they are taught in the right way. They cannot learn them merely from correlations; i.e., merely from observing states of a system. They can learn them in a small number of trials via intervention (Lagnado & Sloman, 2004; Schulz, Kushnir & Gopnik, 2007). Intervention offers several advantages over observation for causal learning: It provides local as opposed to statistical cues; it focuses attention on effective cues; it makes learning active; and, finally, it permits hypothesis testing.

People can also learn causal structure effectively from temporal cues. Delays between cause and effect can be used as guides to trace causal structure (Buehner & May, 2002). In fact, delays are sometimes treated as causal cues even when they are not (Lagnado & Sloman, 2006). This can lead to spurious beliefs about causal relations that do not exist (Fernbach & Sloman, 2009) and illusory beliefs like superstitions (Ono, 1987), illusions of control (Langer, 1975), and illusory correlations (Chapman & Chapman, 1969).

Causal models can also be learned through instruction. This is presumably the primary conduit of causal learning outside immediate experience.

4.5. How do people conceive of cause?

There are two classes of theories about how people answer questions about the "actual" cause of an event (like an accident). One class of theory says that people rely on a counterfactual assessment,

determining what the true cause is by deciding which variables could not have been different in order for the effect to remain unchanged (Lewis, 1973). Alternatively, people can take a causal power or a mechanistic perspective. This involves tracing the invariant property that passed through space and time from cause to effect (Dowe, 2000). With one exception, evidence from the psychological laboratory strongly favors the mechanistic view of causation. People ascribe cause when there is a transmission of a conserved quantity or force (Bullock, 1985; Shultz, 1982; Walsh & Sloman, 2011). They do not ascribe cause when the individual sets the stage for the effect to occur but did not directly contribute to a force that is on the trajectory leading to the outcome. This mechanistic view is consistent with the idea that people mentally simulate a process that involves transfer of a property (like force) from cause to effect (cf. Wolff, 2007).

4.6. Intention as cause

The exception occurs when a cause is intentional. If it is known that someone broke a window intentionally, then people will assert that the person caused the window to break regardless of how they did it (cf. Lombrozo, 2010). They might be on the trajectory connected to the outcome (e.g., they might have thrown a rock at the window) or they might simply have enabled it (e.g., by closing the window to make it vulnerable or even by failing to tell someone else not to throw a rock). It doesn't really matter how the person did it if you believe they wanted the result to occur and they knew it would occur. What counts for ascribing cause is that the person intended it. Presumably the reason that people are more liberal about ascribing intentional causes is that the outcome was likely no matter how it was brought about. There are many ways to break a window; presumably, if one method had been blocked the window breaker would have found another way. This does not hold for unintentional causes since unintentional causes only bring about their effect via the specific causal path that happens to have unfolded. Intentional causes are unique because they have a guiding representation, a will that controls them to achieve a particular end state.

5. Complex causal structures

What the discussion so far suggests is that the first reaction a person would have in the face of a complex causal system is to either use a simplifying heuristic, a linear approximation, or to break the system down by reducing it to a set of cognitively-manageable simpler causal structures.

5.1. Form of representation

The claim that people break systems down into simpler structures is consistent with data suggesting that we tend to construct part-whole hierarchies in a variety of domains. We are clearly easily able to think about, say, biological entities this way. For instance, we can easily perceive the human body as composed of body parts (e.g., the head) which is itself composed of parts (e.g., a face) which can be further broken down into parts (e.g., eyes). You can see the effect of part-whole hierarchical representation in urban planning. Cities and college campuses that are designed by people are quite different than those that evolve naturally (Alexander, 1959). Designed spaces tend to be hierarchical (e.g., a city with distinct commercial, industrial, and residential zones, each with its own police and fire station, etc.). Natural cities are tangled: A particular area or even building can serve many functions (e.g., commercial and residential). In fact, the different functional elements collide constantly. Some people live in largely

commercial areas and commerce is often done in primarily residential areas. Interactions between the different facets of everyday life abound. A resident will often encounter different aspects of city life at the same moment, like when the workers in the factory down the block keep him or her awake at night celebrating at the bar across the street. For this reason, natural cities are much more interesting (and annoying) places to live.

5.2. Information integration

One of peoples' great weaknesses in making probabilistic judgment is our inability to integrate prior beliefs with data appropriately (Bar-Hillel, 1980; Kahneman & Tversky, 1973). If a person's prior beliefs are strong and salient, he or she will tend to give them too much weight both in selecting evidence and in making judgments (Lord, Ross & Lepper, 1979). But prior beliefs, such as knowledge about base rates, will be neglected if they are not part of the story the individual is telling about the case. For instance, it is common knowledge that nonmedical professionals frequently diagnose themselves as having terrible diseases that are also terribly rare. A mild pain or lump can be interpreted according to the worst-case scenario even if that scenario has very low probability in the judge's experience.

5.3. The perception of sequences

People's predictions of the behavior of complex causal systems have been evaluated by asking questions about how they perceive sequences of system outputs. The seminal work in this domain looked at sequences of outcomes in basketball shots (Gilovich, Vallone, & Tversky, 1985), baseball pitches and hits, and other sports performance measures. But how people perceive other kinds of sequences has also been studied, like stock market values over time. The key finding from this research is that people see patterns in randomness; they find structure where there is none. People will classify a sequence produced by a random generating process with independent trials as "streaky." Sequences generated by complex systems sometimes turn out to be random in the sense that outcomes at time t are not predictable from outcomes at previous time points. Sometimes they are not predictable because there are forces causing them not to be (e.g., the efficient market hypothesis; any predictability is a source of profit that can be extracted until the system becomes unpredictable). Sometimes they are not predictable because they are generated by independent events (as in many games of chance). But complex systems can produce random sequences for no apparent reason. Basketball shots, strike outs, sequences of wins, and many more turn out to be sequentially independent.

In all these cases, people see structure where there is not; they see systematicity in the form of periods of strong and weak performance even though trials are in fact independent. And they make predictions accordingly. This has been observed in the laboratory, the casino, among sports fans, and on the trading floor.

6. The causal structure of intuition

People are remarkable in their ability to access and use complex knowledge to generate simple explanations quickly and often with relative ease (Lombrozo, 2006; Sloman, 1994). This requires the ability to deploy a huge amount of knowledge, select what is relevant, and put it together to tell a coherent story, usually a causal story. This tendency people have to make sense of their experience

through narrative governs their lives in small ways – like how they do internet searches (Pirolli & Russell, 2011) – and in large ways – how they make sense of their lives (Bruner, 1991).

There is some evidence that long-term knowledge structures are organized according to causal principles. For instance, Medin, Coley, Storms and Hayes (2003) as well as Kemp and Tenenbaum (2009; also see Shafto et al., 2009) have argued that people use causal structure to make inductive inferences about natural kinds. Specifically, we generalize properties from one category to another based on transfer through a causal mechanism (e.g., germs can get transferred via a member of one species eating a member of another species). Alternatively, we generalize a property from one category to another by evaluating the degree to which the categories are related with respect to a common causal mechanism (e.g., a property is likely to generalize from wolves to German Shepherds by virtue of their common ancestry).

People can even reason about complex systems at multiple levels of abstraction. Such reasoning can be represented by hierarchical Bayesian networks that capture structure at multiple levels (Tenenbaum et al., 2011). Networks like this are the richest representational formalism that has been proposed for human probabilistic reasoning, but their very richness suggests that they may not be psychologically realistic. People's knowledge at lower levels of abstraction tends to be weak, uncertain, and often unavailable. Rozenblit and Keil (2002) have demonstrated that people believe they know much more about objects and systems than in fact they do ("the illusion of explanatory depth"). For instance, men tend to think that they can describe the mechanism by which a toilet works, but when queried most of them demonstrate little understanding. In fact, the mere query causes many people to reduce their rating of how well they understand the object. The evidence suggests that people have fairly abstract knowledge about how things work but few of the details.

There is in fact a variety of evidence that intuitive causal knowledge does not reflect a thorough understanding of the underlying complex system. Indirect evidence for this claim comes from the fact that there are limits on our ability to explain our own attitudes. A powerful example of this is moral dumbfounding (Haidt, 2001) in which agents cannot justify their moral attitudes. For instance, many people consider it immoral for siblings to sleep with each other. And they continue to think so concerning a particular case in which all the available reasons for objecting have been undermined (e.g., by assuming pregnancy is impossible, no coercion, etc.). In other words, an incest taboo survives even when all available justifications for it do not. More generally, we have many attitudes and beliefs that we cannot explain and justify. The illusion of explanatory depth suggests that this extends to causal beliefs. In fact, there is evidence that the explanations people do provide for their beliefs are really post-hoc justifications that do not derive from the same cognitive processes that produced the beliefs (Nisbett & Wilson, 1977). In fact, asking people to explain why they think they will perform some action in the future before making a prediction about their own action can destroy what would otherwise have been an accurate prediction (e.g., Wilson et al., 1993). For instance, asking students to analyze their reasons for prefering certain college courses before course selection caused them to choose courses that corresponded less well with expert opinion than those who did not analyze their reasons (Wilson & Schooler, 1991).

Domain experts clearly have rich causal stories to tell. As we saw above, they use these causal stories to make clinical judgments (Dawes, Faust, Meehl, 1989). They seem to emerge from rich inferential knowledge bases in the form of patterns that support inference. Experts perceive complex patterns that are invisible to novices. This has been shown experimentally in a number of domains including chess (Chase & Simon, 1973) and dermatology (Brooks, Norman, & Allen, 1991). In clinical domains lacking good models, experts also perceive patterns. But those patterns tend not to have huge amounts of predictive power, less than that offered by simple linear combinations (Dawes, 1979).

6.1. Summary of work on intuitive causal models

What these data suggest is that our long-term knowledge base is organized according to causal principles. The information we extract from it in the form of patterns and stories does have causal structure. Although that causal structure carries a lot of information, it can also be biased because the information that is extracted is determined by the query addressed to the knowledge base, and that query can be shaped by ideology, neglect of alternative causes, recent events, the need to justify, and possibly other factors.

7. Motivated reasoning

One issue that we will not give a full treatment to is the role of emotion and motivation in reasoning. Emotion is an important contributor to good decision making by harnessing attention and focusing it on issues of greatest relevance (Damasio, 1994). Emotional barriers can also inhibit the quality of reasoning in two ways: For one, they can distract us making it harder to focus attention where it would best serve us. This depends on having the capapcity to focus and concentrate in the first place. Beilock and Carr (2005) gave students a mathematics lesson and then tested them and found that students with lower working memory capacity were not affected by stress but those with higher working memory capacity were. Presumably the students with lower capacity were not affected because they were already distracted.

A second way emotion that emotion can bias us is by influencing our goals, leading us to reason in favor of a pre-determined conclusion rather than to draw the conclusions dictated by logic (Kunda, 1990). This kind of effect has been observed in the domain of causal reasoning. Quattrone and Tversky (1984) gave people a test of their tolerance for pain under two conditions. In one, they were told that people can tolerate pain better if they have a good type of heart. In the second, participants were told that those with a weaker type of heart had higher tolerance. Not surprisingly, the first group showed greater tolerance than the second group. This entails a form of self-deception in that the act of manipulating their tolerance for pain made it nondiagnostic of their heart type. To the extent they were manipulating their tolerance via an act of will, their hearts could not have been determining their response, thus making their tolerance uninformative about their heart type.

Sloman, Fernbach, and Hagmayer (2010) refer to cases like this in which people believe they have less control over their actions than they do as “diagnostic self-deception”. They also posit a different kind of self-deception that underlies addiction that they call “interventional self-deception”. In such cases, people believe they have more control over their actions than in fact they do, just as teenagers who smoke tend to believe that they can quit at any time. Empirically, we showed that people will not use an accurate causal representation but instead will apply self-serving frames leading to self-deception, but only given sufficient ambiguity in the environment.

8. Conclusion

From the perspective of a person with uncertain knowledge, sociotechnical systems that are sufficiently complex are hard to predict because future states are inevitably affected by variables whose state is unknown. In fact the very existence of the variables may be unknown. When the causal structure of a system is unknown, prediction with a Causal Bayes net is difficult because it requires calculating an expectation over all possible causal structures based on their probabilities, and those probabilities are

also likely to be unknown. Analogous difficulties exist for other formal predictive systems. This source of difficulty is inherent to complex systems.

The nature of human cognition brings several additional obstacles to the table. People tend to use simplifying heuristics and linear approximations whose value is limited when systems are complex and highly nonlinear. When we think causally, we tend to think in terms of small, local, acyclic structures that do not capture much of the behavior of a complex, dynamic system. We tend to make predictions by mentally simulating how causes lead to their effects and this leads us to neglect alternative causes of events. We may even use this basic cognitive simulation facility when we are making diagnoses and it is possible that this leads us to make different kinds of errors in that domain, like neglecting enabling and disabling conditions. These habits of mind are often useful but lead to systematic error, especially in the face of complexity. We have seen that people have trouble perceiving randomness; instead they see structure that is not there.

In sum, we cannot assume optimality to model the human contributions to a complex system, or to explain how a complex system works to society at large. People do not think in ways that correspond to how complex systems actually work. But the good news is that we are beginning understand how people do think, and this knowledge can be leveraged to build models of people that may not be optimal, but at least they will be accurate.

Steven Sloman is a Professor in the Department of Cognitive, Linguistic, and Psychological Sciences at Brown University. He received his B.Sc. from the University of Toronto in 1986 and his Ph.D. from Stanford University in 1990, both in psychology. He is a computationally-oriented experimental cognitive scientist whose work concerns higher-order aspects of cognition, including causal reasoning, decision making, judgment, and categorization. Sloman's work focuses on two themes, the logic of human inference and choice and the architecture of mind. His published work includes two books, most recently Causal Models: How We Think About the World and Its Alternatives (2005). He is also known as a purveyor of dual-system theories, having argued that inferential processes come in two types, an intuitive one and a deliberative one. Sloman is currently chief associate editor of the journal Cognition.

Philip Fernbach is an assistant professor in the Leeds School of Business at the University of Colorado, Boulder. His research interests span many areas of high-level cognition including judgment and decision-making, planning and goal pursuit, reasoning, mental state ascription and moral judgment. Most of his work is inspired by causal model theory, the idea that people's inductive inferences are based on knowledge of the world, knowledge that is represented in terms of causal structure. His research has appeared in the premier journals in psychology including Psychological Science, The Journal of Experimental Psychology and Cognition and has been profiled in places like ABCnews and the Boston Globe. He received his Ph.D. in cognitive science from Brown University in 2010 and a B.A. in philosophy from Williams College in 2001.

References

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Alter, A. L. and Oppenheimer, D. M. (2009), Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219-35.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-609.

Bes, B., Sloman, S. A., Lucas, C. G. and Raufaste, E. (2010). Non-Bayesian inference: Causal structure trumps correlation. Submitted for publication.

Beilock, S. L. and Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16*, 101-105.

Blaisdell, A. P., Sawa, K., Leising, K. J. and Waldmann, M. R. (2006, February 17). Causal reasoning in rats. *Science, 311*, 1020-1022.

Bruner, J. S. (1991). The narrative construction of reality. *Critical Inquiry, 18*(1), 1-21.

Brunswik, E. (1955), Representative design and probabilistic theory in functional psychology, *Psychological Review, 62*, 193-217.

Buehner, M. J. and May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning, 8*(4), 269-295.

Chapman, L. J. and Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnorma Psychology, 74*, 271-280.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R.M., Faust, D., Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Dowe, P. (2000). *Physical causation*. Cambridge, England: Cambridge University Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. and Plunkett, K. (1996). Rethinking Innateness. A Connectionist Perspective on Development. Cambridge, MA: MIT Press.

Fernbach, P. M. and Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 678-693.

Fernbach, P. M., Darlow A. and Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21*(3), 329-336.

Fernbach, P. M., Darlow, A. and Sloman, S. A. (2011a). When good evidence goes bad. The weak evidence effect in judgment and decision-making. *Cognition, 119*, 459-467.

Fernbach, P. M., Darlow A. and Sloman, S. A. (2011b). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140*(2), 168-185.

Fernbach, P. M. and Rehder, B. (2011). Cognitive shortcuts in causal inference. *Manuscript submitted for publication*.

Gentner D. and Stevens A. L. (2003) (eds.) *Mental Models*. Hillsdale, NJ: Erlbaum.

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

Gilovich, T., Vallone, R. and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

Gilovich, T. and Griffin, D. (2002). Heuristics and biases: Then and now. In T. Gilovich, D. W. Griffin. and D. Kahneman (Eds). *Heuristics and biases: The psychology of intuitive judgment* (pp. 230-249). Cambridge, England: Cambridge University Press.

Hagmayer, Y. and Sloman, S. A. (2009). Decision makers conceive of their choice as intervention. *Journal of Experimental Psychology: General, 138*, 22-38.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814-834.

Hegarty, M. (2004). Mechanical reasoning as mental simulation. *TRENDS in Cognitive Sciences, 8*, 280-285.

Hogarth, R. M. and Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge: Bradford Books.

Jenkins, H. M. and Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79*, 1-17.

Jones, E. and Nisbett R. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins and B. Weiner (Eds), *Attribution: Perceiving the causes of behavior* (pp. 121-135). Morristown, NJ: General Learning Press.

Kahneman, D. (2011). Thinking fast and slow. New York: Farrar, Straus and Giroux.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kahneman, D. and Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201-208). New York, NY: Cambridge University Press.

Kaufmann, Stefan. 2009. Conditionals right and left: Probabilities for the whole family. *Journal of Philosophical Logic, 38*(1):1-53.

Kelley, H. H. (1972a). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins and B. Weiner (Eds), *Attribution: Perceiving the causes of behavior* (pp. 11-26). Morristown, NJ: General Learning Press.

Kemp, C. and Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*, 20-58.

Kun, A., Murray, J. and Sredl, T. (1980). Misuses of the multiple sufficient causal scheme as a model of naive attributions: A case of mistaken identity. *Developmental Psychology, 16*, 13-22.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3): 480-498.

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

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

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*, 311-328.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*, 556-567.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464-470.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*, 303-332.

Lord, C. G., Ross, L. and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098-2109.

Medin, D. L., Coley, J. D., Storms, G. and Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review, 10*, 517-532.

Morris, M. W. and Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*, 331-355.

Nisbett, R. and Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

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

Park, J. H. (2011) When does screening off hold in causal reasoning? *Manuscript in preparation*.

Pearl, J. (2000). *Causality*. Cambridge, England: Cambridge University Press.

Pennington, N. and Hastie, R. (1993). Reasoning in explanation-based decision-making. *Cognition, 49*, 123-163.

Pirolli, P. Russell, D. M. (2011). Special issue on "Sensemaking." *Human-Computer Interaction Journal, 26*, n. 1 and 2.

Quattrone, G. A. and Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and the voter's illusion. *Journal of Personality and Social Psychology, 46*, 237-248.

Rehder, B. and Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*(3), 264-314.

Rips, L. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science, 34*, 175-221.

Rozenblit, L. and Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 92*, 1-42.

Schulz, L. E., Kushnir, T. and Gopnik, A. (2007). Learning from doing: Interventions and causal inference. In A. Gopnik and L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86-100). Oxford, UK: Oxford University Press.

Shafto, P., Kemp, C., Baraff Bonawitz, E., Coley, J. D. and Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109*, 175-192.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*, 1-51.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52*, 1-21.

Sloman, S. A., Fernbach, P. M. and Hagmayer, Y. (2010). Self deception requires vagueness. *Cognition, 115*(2), 268-281.

Sloman, S. A. and Lagnado, D. (2005). Do we "do"? *Cognitive Science, 29*, 5-39.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, prediction and search*. New York, NY: Springer–Verlag.

Teigen, K. H. (2005). The proximity heuristic in judgments of accident probabilities. *British Journal of Psychology, 96*(4), 423-440.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science, 331*(6022), 1279-1285.

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology 5*, 207-232.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315.

Waldmann, M. R. and Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Walsh, C. and Sloman, S. A. (2008). Updating beliefs with causal models: violations of screening off. In M. A. Gluck, J. R. Anderson and S. M. Kosslyn (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (pp. 345-357). Mahwah, NJ: Erlbaum.

Walsh, C. R. and Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind and Language*. In press.

Wells, G. L. and Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*, 161-169.

Wilson, T. D., Lisle, D., Schooler, J. W., Hodges, S. D., Klaaren, K. J. and LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19*, 331-339.

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

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*, 82-111.