Causal Mechanisms

Samuel G. B. Johnson

Woo-kyoung Ahn

Department of Psychology, Yale University

Corresponding author:

Samuel Johnson
2 Hillhouse Ave.
New Haven, CT 06520
samuel.johnson@yale.edu
262-758-9744

In press in *Oxford Handbook of Causal Reasoning* (Michael R. Waldmann, Ed.)

Note. This is a prepublication copy of this chapter. The final published version may difer.

Abstract

This chapter reviews empirical and theoretical results concerning knowledge of *causal mechanisms*—beliefs about how and why events are causally linked. First, we review the effects of mechanism knowledge, showing that mechanism knowledge can trump other cues to causality (including covariation evidence and temporal cues) and structural constraints (the Markov condition), and that mechanisms play a key role in various forms of inductive inference. Second, we examine several theories of how mechanisms are mentally represented—as associations, forces or powers, icons, abstract placeholders, networks, or schemas—and the empirical evidence bearing on each theory. Finally, we describe ways that people acquire mechanism knowledge, discussing the contributions from statistical induction, testimony, reasoning, and perception. For each of these topics, we highlight key open questions for future research.

Keywords: Causal mechanisms, causal representation, causal learning, causal chains, induction

Introduction

Our causal knowledge not only includes beliefs about which events are caused by other events, but also an understanding of how and why those events are related. For instance, when a soprano hits an extremely high note, the sound can break a wine glass due to the high frequency of the sound waves. Although people may not know the detailed mechanisms underlying this relationship (Rozenblit & Keil, 2002), people believe that *some* mechanism transmits a force from the cause to the effect (White, 1989). Likewise, people believe in causal mechanisms underlying interpersonal relations (see Hilton, this volume). When Romeo calls to the balcony, Juliet comes, and she does so because of her love. When Claudius murders the king, Hamlet seeks revenge, because Hamlet is filled with rage. We use mechanisms to reason about topics as grand as science (Koslowski, 1996) and morality (Cushman, 2008; see Lagnado & Gerstenberg, this volume); and domains as diverse as collision events (Gerstenberg & Tenenbaum, this volume; White, this volume) and psychopathology (Ahn, Kim, & Lebowitz, this volume). Causal mechanisms pervade our cognition through and through.

Indeed, when a person tries to determine the cause of an event, understanding the underlying causal mechanism appears to be the primary concern. For instance, when attempting to identify the cause of "John had an accident on Route 7 yesterday," participants in Ahn, Kalish, Medin, and Gelman (1995) usually asked questions aimed at testing possible mechanisms (e.g., "Was John drunk" or "Was there a mechanical problem with the car?") rather than which factor was responsible for the effect (e.g., "Was there something special about John?" or "Did other people also have a traffic accident last night?").

In this chapter, we describe the state of current research on mechanism knowledge. After defining terms, we review the effects of mechanism knowledge. We summarize studies showing

(1) that mechanism knowledge can trump other important cues to causality, and (2) that mechanism knowledge is critical for inductive inference. Next, we examine how mechanisms might be mentally represented, and summarize the empirical evidence bearing on each of several approaches. We then turn to how mechanisms are learned, parsing the contributions from statistical induction, testimony, reasoning, and perception. For each of these broad topics, we discuss potential avenues of future research.

What is a Causal Mechanism?

A causal mechanism is generally defined as a (i) system of physical parts or abstract variables that (ii) causally interact in systematically predictable ways so that their operation can be generalized to new situations (e.g., Glennan, 1996; Machamer, Darden, & Craver, 2000). We use the term *mechanism knowledge* to refer to a mental representation of such a system.

Mechanism knowledge is critical in cognition because we use it to understand other causal relations (Ahn & Kalish, 2000). Thus, we are motivated to seek out the mechanisms that *underlie* a causal relationship. The mechanism underlying the relation "*X* caused *Y*" (e.g., a soprano's singing caused a wine glass to break) will involve constructs *other than X and Y* (e.g., high frequency of the voice), but which can connect those events together. For this reason, mechanisms have a close relationship to explanations (Lombrozo, 2010, Lombrozo & Vasilyeva, this volume). For instance, the causal relation "Mary was talking on her cell phone and crashed into a truck" can be explained through its underlying mechanism, "Mary was distracted and didn't see the red light." However, because causal knowledge is organized hierarchically (Johnson & Keil, 2014; Simon, 1996), this entire causal system could be embedded into a larger system such that more specific events might act as mechanisms underlying more general events. That is, "Mary was talking on her cell phone and crashed into a truck" might be a mechanism

underlying "Mary's driving caused a traffic accident," which in turn might be a mechanism underlying "Mary caused delays on I-95," and so on. Thus, mechanism knowledge is not merely a belief about what *caused* some event, but a belief about *how* or *why* that event was brought about by its cause, which can itself be explained in terms of another underlying mechanism, *ad infinitum*. Although we adopt this understanding of mechanism as a working definition, other factors such as the organization of memory appear to play a role in how mechanism knowledge is used and in what counts as a mechanism (Johnson & Ahn, 2015). We discuss some of these factors later in this chapter (see "Representing Mechanism Knowledge").

The term 'mechanism' has also been used in several other ways in the literature, which are somewhat different from our use. First, the term 'mechanistic explanation' is used to refer to backward-looking explanations (e.g., the knife is sharp because Mark filed it), as opposed to forward-looking, teleological explanations (the knife is sharp because it is for cutting; Lombrozo, 2010). However, this distinction does not map onto our sense of mechanism, because teleological explanations can often be recast in mechanistic terms, in terms of causally interacting variables (e.g., the knife is sharp because human agents wanted to fashion a sharp object, and forging a sharp piece of metal was the best way to accomplish this goal; Lombrozo & Carey, 2006).

Second, some have argued that our knowledge of mechanisms underlying two causally related events, say *A* and *B*, includes not only the belief that there is a system of causally related variables mediating the relationship between *A* and *B* (a 'mechanism' as defined in the current chapter), but also an assumption that a force or causal power is transmitted from A to B (Ahn & Kalish, 2000; White, 1989). This is an independent issue because knowledge about a system of causally interconnected parts does not have to involve the notion of causal power or force. In fact, many of studies reviewed in this chapter demonstrating effects of mechanism knowledge did not

test whether the assumptions of causal force are required to obtain such effects. In this chapter, we separate these two issues when defining mechanism knowledge. Thus, our discussion of the effects of mechanism knowledge does not take a position on the debate concerning causal force, and our discussions of how people represent and learn mechanisms do not beg the question against statistical theories.

Using Mechanism Knowledge

A major purpose of high-level cognition is *inductive inference*—predicting the unknown from the known. Here, we argue that mechanism knowledge plays a critical role in people's inductive capacities. We describe studies on how mechanism knowledge is used in a variety of inductive tasks, including causal inference, category formation, category-based induction, and probability judgment.

Mechanisms and Causal Inference

David Hume (1977/1748) identified two cues as critical to identifying causal relationships—covariation (the cause and effect occurring on the same occasions more often than would be expected by chance) and temporal contiguity (the cause and effect occurring close together in time). Both of these factors have received considerable empirical attention in recent years, and it has become increasingly clear that neither of these cues acts alone, but rather in conjunction with prior knowledge of causal mechanisms. In this section, we first describe how mechanism knowledge influences the interpretation of covariation information. We then describe how mechanism knowledge can result in violations of the Causal Markov Condition, a key assumption to modern Bayesian approaches to causal inference. Finally, we review evidence that even the seemingly straightforward cue of temporal contiguity is influenced in a top-down manner by mechanism knowledge.

Covariation. Scientists must test their hypotheses using statistical inference. To know whether a medical treatment really works, or a genetic mutation really has a certain effect, or a psychological principle really applies, one must test whether the cause and effect are statistically associated. This observation leads to the plausible conjecture that laypeople's everyday causal reasoning also depends on an ability to test for covariation between cause and effect.

But consider the following (real) research finding from medical science (Focht, Spicer, & Fairchok, 2002): Placing duct tape over a wart made it disappear in 85% of the cases (compared to 60% of cases receiving more traditional cryotherapy). Despite the study's experimental manipulation and statistically significant effect, people may still be doubtful that duct tape can remove warts because they cannot think of a plausible mechanism underlying the causal relationship. In fact, the researchers supplied a mechanism: the duct tape irritates the skin, which in turn stimulates an immune system response, which in turn wipes out the viral infection that had caused the wart in the first place. Given this mechanism information, people would be far likelier to believe this causal link. Thus, even statistically compelling covariation obtained through experimental manipulation may not be taken as evidence for a causal link in the absence of a plausible underlying mechanism.

However, in this example, it could be that the mechanism is supplying 'covert' covariation information—for example, the mechanism implies covariation between duct tape and irritation, irritation and immune response, and immune response and wart recovery, and could have thereby conveyed stronger covariation between duct tape and wart recovery. In that case, one might argue that there is nothing special about mechanism information other than conveying covariation. To empirically demonstrate that mechanism information bolsters causal inferences above and beyond the covariation implied by the mechanism, Ahn et al. (1995, Experiment 4)

asked a group of participants to rate the strength of the covariation implied by sentences like "John does not know how to drive" for "John had a traffic accident." They then asked a new group of participants to make causal attributions for the effect (e.g., the accident), given either the mechanism (e.g., John does not know how to drive) or its equivalent covariation (e.g., John is much more likely to have a traffic accident than other people are), as rated by the first group of participants. Participants were much more inclined to attribute the accident to the target cause when given the underlying mechanism, showing that mechanism information has an effect that goes beyond covariation.

More generally, the interpretation of covariation data is strongly influenced by mechanism knowledge. For example, learning about a covariation between a cause and effect has a stronger effect on the judged probability of a causal relationship when there is a plausible mechanism underlying the cause and effect (e.g., severed break lines and a car accident) than when there is not (e.g., a flat tire and a car failing to start; Fugelsang & Thompson, 2000). Similarly, both scientists and laypeople are more likely to discount data inconsistent with an existing causal theory, relative to data consistent with the theory (Fugelsang, Stein, Green, & Dunbar, 2004). Finally, people are more likely to condition on a potential alternative cause when interpreting trial-by-trial contingency data, if they are told about the mechanism by which the alternative cause operates (Spellman, Price, & Logan, 2001). These effects show that not only does mechanism information do something beyond covariation, but that it even constrains the way that covariation is used.

Structural constraints. Patterns of covariation between variables can be combined into larger patterns of causal dependency, represented as Bayesian networks (Pearl, 2000; Rottman & Hastie, 2014; Rottman, this volume). For example, if a covariation is known to exist between

smoking cigarettes (*A*) and impairment of lung function (*B*), and another is known to exist between smoking cigarettes (*A*) and financial burden (*C*), this can be represented as a causal network with an arrow from *A* to *B* and an arrow from *A* to *C* (a *common cause* structure). But of course, all of these events also have causes and consequences—social pressure causes cigarette smoking, impairment of lung function causes less frequent exercise, financial burden causes marital stress; and so on, *ad infinitum*. If we had to take into account all of these variables to make predictions about any of them (say, *B*), then we would never be able to use causal knowledge to do anything. The world is replete with too much information for cognition without constraints.

The key computational constraint posited by Bayesian network theories of causation is the *Causal Markov Condition* (also known as 'screening off'; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). This assumption allows the reasoner to ignore the vast majority of potential variables—to assume that the probability distribution of a given variable is independent of all other variables except its direct effects, conditional on its causes. For example, the Markov condition tells us, given the causal structure described above for smoking, that if we know that Lisa smokes (*A*), knowing about her lung function (*B*) doesn't tell us anything about her potential financial burden (*C*), and vice versa. Because the Markov Condition is what allows reasoners to ignore irrelevant variables (here, we can predict *B* without knowing about *C* or any of the causes of *A*), it is crucial for inference on Bayesian networks.

Alas, people often violate the Markov Condition. Although there appear to be a number of factors at play in these violations, including essentialist (Rehder & Burnett, 2005) and associationist (Rehder, 2014) thinking, one critical factor is mechanism knowledge (Park & Sloman, 2013, 2014). In common cause structures such as the smoking example above (smoking

leading to lung impairment and financial burden) where each causal link relies on a different mechanism, people do tend to obey the Markov Condition. That is, when asked to judge the probability of lung impairment given that a person smokes, this judgment is the same as when asked to judge the probability of lung impairment given that a person smokes and has a financial burden. But when the links rely on the *same* mechanism (e.g., smoking leading to lung impairment and to blood vessel damage), people robustly violate the Markov condition. When asked to judge the probability of lung impairment given that a person smokes, this judgment is lower than when asked to judge the probability of lung impairment given that a person smokes and has blood vessel damage.

This effect is thought to occur because participants use mechanism information to elaborate on the causal structure, interpolating the underlying mechanism into the causal graph (Park & Sloman, 2013). So, when the link between *A* and *B* depends on a different mechanism than the link between *A* and *C*, the resulting structure would involve two branches emanating from *A*, namely $A \rightarrow M_1 \rightarrow B$ and $A \rightarrow M_2 \rightarrow C$. In Lisa's case, cellular damage might be the mechanism mediating smoking and lung impairment, but cigarette expenditures would be the mechanism mediating smoking and financial burden. Thus, knowing about *C* (Lisa's financial burden) triggers an inference about M_2 (cigarette expenditures), but this knowledge has no effect on *B* (lung impairment) given that *A* (smoking) is known—the Markov condition is respected. But when the link between *A* and *B* depends on the same mechanism as the link between *A* and *C*, the resulting structure would be a link from *A* to M_1, and then from M_1 to *B* and to *C*—so, in effect, the mechanism M_1 is the common cause, rather than *A*. That is, cellular damage might be the mechanism mediating the relationship between smoking and lung impairment *and* the relationship between smoking and blood vessel damage. Thus, knowing about *C* (blood vessel

damage) triggers an inference about M_1 (cellular damage), and this knowledge has an effect on *B* (lung impairment) even if *A* (smoking) is known. Mechanism knowledge therefore not only affects the interpretation of covariation information, but also the very computational principles used to make inferences over systems of variables.

Temporal cues. According to the principle of *temporal contiguity*, two events are more likely to be causally connected if they occur close together in time. This idea has considerable empirical support (e.g., Lagnado & Sloman, 2006; Michotte, 1963/1946), and at least in some contexts, temporal contiguity appears to be used more readily than covariation in learning causal relations (Rottman & Keil, 2012; White, 2006). The use of temporal contiguity was long taken as a triumph for associationist theories of causal inference (Shanks, Pearson, & Dickinson, 1989), because longer temporal delays are associated with weaker associations in associationist learning models.

Yet, people's use of temporal cues appears to be more nuanced. People are able to associate causes and effects that are very distant in time (Einhorn & Hogarth, 1986). For example, a long temporal gap intervenes between sex and birth, between smoking and cancer, between work and paycheck, and between murder and prison. Why is it that the long temporal gaps between these events do not prevent us from noticing these causal links?

A series of papers by Buehner and colleagues documented top-down influences of causal knowledge on the use temporal contiguity (see Buehner, this volume). When participants expect a delay between cause and effect, longer delays have a markedly smaller deleterious effect on causal inference (Buehner & May, 2002, 2003), suggesting some knowledge mediation. In fact, when temporal delay is de-confounded with contingency, the effect of temporal delay can be eliminated altogether by instructions that induce the expectation of delay (Buehner & May,

2004). Most dramatically, some experiments used unseen physical causal mechanisms, which participants would believe to take a relatively short time to operate (a ball rolling down a steep ramp, hidden from view) or a long time to operate (a ball rolling down a shallow ramp). Under such circumstances, causal judgments were *facilitated* by longer delays between cause and effect, when the mechanism was one which would take a relatively long time to operate (Buehner & McGregor, 2006). Although older (9- to 10-year-old) children can integrate such mechanism cues with temporal information, younger (4- to 8-year-old) children continued to be swayed by temporal contiguity, suggesting that the relative priority of causal cues undergoes development (Schlottmann, 1999). Thus, when people can apply a mechanism to a putative causal relationship, they adjust their expectations about temporal delay so as to fit their knowledge of that mechanism.

Mechanisms and Induction

The *raison d'être* for high-level cognition in general, and for causal inference in particular, is to infer the unknown from the known—to make predictions that will usefully serve the organism through *inductive inference* (Murphy, 2002; Rehder, this volume a, b). In this section, we give several examples of ways that mechanism knowledge is critical to inductive inference.

Categories are a prototypical cognitive structure that exists to support inductive inference. We group together entities with similar known properties, because those entities are likely to also share similar unknown properties (Murphy, 2002). Mechanism knowledge influences which categories we use. In a study by Hagmayer, Meder, von Sydow, and Waldmann (2011), participants learned the contingency between molecules and cell death. Molecules varied in size (large or small) and color (white or grey). While large white (11) molecules always led to cell

death and small grey (00) molecules never did, small white (01) and large grey (10) ones led to cell death 50% of the time. That is, 01 and 10 were equally predictive of cell death. However, prior to this contingency learning, some participants learned that molecule color was caused by a genetic mutation. Participants used this prior causal history to categorize small white molecules (01) with large white (11) molecules, which always resulted in cell death. Consequently, these participants judged that small white molecules (01) were much more likely to result in cell death than large grey molecules (10), even though they observed both probabilities to be 50%. The opposite pattern was obtained when participants learned that genetic mutation caused molecules to be large.

Critically, this effect of prior categorization on subsequent causal learning depended on the type of underlying mechanism. Note that most people would agree that genetic mutations affect deeper features of molecules, which not only affects surface features such as color of molecules, but also can affect likelihood of cell death. Thus, the initial category learning based on the cover story involving genetic mutations provided a mechanism, which could affect later causal judgments involving cell death. In a subsequent experiment, however, the cover story used for category learning provided an incoherent mechanism. Participants learned that the variations in color (or size) were due to atmospheric pressure, which would be viewed as affecting only the surface features. Despite identical learning situations, participants provided with mechanism information that were relevant only to surface features did not distinguish between 10 and 01 in their causal judgments; their judgments stayed close to 50%. Thus, Hagmayer et al. (2011) showed that prior learning of categorization affects subsequent causal judgments only when the categorization involves mechanisms that would be relevant to the content of the causal judgments (see also Waldmann & Hagmayer, 2006 for related results).

More generally, people are likely to induce and use categories that are *coherent* (Murphy & Allopenna, 1994; Rehder & Hastie, 2004; Rehder & Ross, 2001). A category is coherent to the extent that its features 'go together', given the reasoner's prior causal theories (Murphy & Medin, 1985). For example, "lives in water, eats fish, has many offspring, and is small" is a coherent category, because one can think of a causal mechanism that unifies these features, supplying the necessary mechanism knowledge; in contrast, "lives in water, eats wheat, has a flat end, and is used for stabbing bugs" is an incoherent category because it is difficult to supply mechanisms that could unify these features in a single causal theory (Murphy & Wisniewski, 1989). Categories based on a coherent mechanism are easier to learn (Murphy, 2002), more likely to support the extension of properties to new members (Rehder & Hastie, 2004), and require fewer members possessing a given property to do so (Patalano & Ross, 2007).

Mechanism knowledge also influences *category-based induction*, or the likelihood of extending features from one category to another (see Heit, 2000 for a review). If the mechanism explaining why the premise category has a property is the same as the mechanism explaining why the conclusion category might have the property, then participants tend to rate the conclusion category as very likely having that property (Sloman, 1994). For example, participants found the following argument highly convincing:

Hyundais have tariffs applied to them; therefore,

Porsches have tariffs applied to them.

That is, the reason that Hyundais have tariffs applied to them is because they are foreign cars, which would also explain why Porsches have tariffs applied to them. So, the premise in this case strongly supports the conclusion. In contrast, one may discount the likelihood of a conclusion when the premise and conclusion rely on different mechanisms, such as:

Hyundais are usually purchased by people 25 years old and younger; therefore,

Porsches are usually purchased by people 25 years old and younger.

In this case, the reason that Hyundais are purchased by young people (that Hyundais are inexpensive and young people do not have good credit) does not apply to Porsches (which might be purchased by young people because young people like fast cars). Because the premise introduces an alternative explanation for the property, people tend to rate the probability of the conclusion about Porsches *lower* when the premise about Hyundais is given, compared to when it is not given—an instance of the *discounting* or *explaining-away effect* (Kelley, 1973). These results show that mechanism knowledge can moderate the likelihood of accepting an explanation in the presence of another explanation.

Ahn and Bailenson (1996) further examined the role of mechanism knowledge in the discounting and conjunction effects. In the discounting effect (Kelley, 1973), people rate the probability P(*B*) of one explanation higher than its conditional probability given another competing explanation, P(*B*|*A*). In the conjunction effect (Tversky & Kahneman, 1983), people rate the probability of a conjunctive explanation, P(*A*&*B*), higher than its individual constituents such as P(*A*). The two effects may appear contradictory because the discounting effect seems to imply that one explanation is better than two, whereas the conjunction effect seems to imply that two explanations are better than one. Yet, Ahn and Bailenson (1996) showed that both phenomena turn on mechanism-based reasoning, and can occur simultaneously with identical events. For example, consider the task of explaining why Kim had a traffic accident. Further suppose that a reasoner learns that Kim is nearsighted. Given this explanation, a reasoner can imagine Kim having a traffic accident due to her nearsightedness. Note that to accept this explanation, one has to imagine that Kim's nearsightedness is severe enough to cause a traffic

accident even under normal circumstances. Once such a mechanism is established, another explanation, “There was a severe storm,” would be seen as less likely because Kim’s nearsightedness is already a sufficient cause for a traffic accident. Thus, the second cause would be discounted. However, consider a different situation where both explanations are presented as being tentative and to be evaluated simultaneously. Thus, one is to judge the likelihood that Kim had a traffic accident because she is nearsighted and there was a severe storm. In this case, a reasoner can portray a slightly different, yet coherent mechanism where Kim’s (somewhat) poor vision coupled with poor visibility caused by a storm would have led to a traffic accident. Due to this coherent mechanism, the reasoner would be willing to accept the conjunctive explanation as highly likely—even as more likely than either of its conjuncts individually. That is, the discounting effect occurs because a reasoner settles in on a mechanism that excludes a second explanation, whereas the conjunction effect occurs because a reasoner can construct a coherent mechanism that can incorporate both explanations.

In addition to demonstrating simultaneous conjunction and discounting effects, Ahn and Bailenson (1996) further showed that these effects do not occur when explanations are purely covariation-based—that is, when the explanations indicate positive covariation between a potential cause and effect without suggesting any underlying mechanism mediating their relationship. For instance, the explanations “Kim is more likely to have traffic accidents than other people are” and “traffic accidents were more likely to occur last night than on other nights” resulted in neither conjunction nor discounting effects. This pattern of results indicates that both discounting and conjunction effects are species of mechanism-based reasoning.

Open Questions

These studies demonstrate a variety of ways that mechanism knowledge pervades our inductive capacities, but mechanism knowledge could affect induction in yet other ways. Beyond covariation, structural constraints, and temporal cues, might other cues to causality be affected by the nature of the underlying mechanisms? For instance, might the results of interventions be interpreted differently given different mechanisms? Might mechanism knowledge modulate the relative importance of these various cues to causality?

There are also open questions about how mechanisms are used in induction. Given the tight link between mechanisms and explanation, what role might mechanisms play in inference to the best explanation, or abductive inference (Lipton, 2004; Lombrozo, 2012)? To what extent do different sorts of inductive problems (Kemp & Jern, 2013) lend themselves more to mechanism-based versus probability-based causal reasoning (see also Lombrozo, 2010)? Are there individual differences in the use of mechanisms? For instance, given that mechanisms underlie surface events, could people who are more intolerant of ambiguity or more in need of cognitive closure be more motivated to seek them out? Could people who are high in creativity be more capable of generating them, and more affected by them as a result? Finally, although we could in principle keep on asking "why" questions perpetually, we eventually settle for a given level of detail as adequate. What determines this optimal level of mechanistic explanation?

Representing Causal Mechanisms

In the previous section, we described several of the cognitive processes that use mechanism knowledge. Here, we ask how mechanism knowledge is *mentally represented* (Markman, 1999). That is, what information do we store about mechanisms, and how do different mechanisms relate to one another in memory? We consider six possible representational formats—associations, forces or powers, icons, abstract placeholders, networks, and schemas.

Associations

According to associationist theories of causality, learning about causal relationships is equivalent to learning associations between causes and effects, using domain-general learning mechanisms that are evolutionarily ancient and used in other areas of causation (Shanks, 1987; Le Pelley, Griffiths, & Beesley, this volume). Thus, causal relations (including mechanism knowledge) would be represented as an association between two classes of events, akin to the stored result of a statistical significance test, so that one event would lead to the expectation of the other. This view is theoretically economical, in that associative learning is well-established and well-understood in other domains and in animal models. Further, associative learning can explain many effects in trial-by-trial causal learning experiments, including effects of contingency (Shanks, 1987) and delay (Shanks, Pearson, & Dickinson, 1989).

However, hard times have fallen on purely associative theories of causation. Because these theories generally do not distinguish between the role of cause and effect, they have difficulty accounting for asymmetries in predictive and diagnostic causal learning (Waldmann, 2000; Waldmann & Holyoak, 1992). Further, these theories predict a monotonic decline in associative strength with a delay between cause and effect, yet this decline can be eliminated or even reversed with appropriate mechanism knowledge (Buehner & May, 2004; Buehner & McGregor, 2006). Although associative processes are likely to play some role in causal reasoning and learning (e.g., Rehder, 2014), causal learning appears to go beyond mere association.

There are also problems with associations as representations of mechanism knowledge. One straightforward way of representing mechanism knowledge using associations is to represent causal relations among sub-parts or intermediate steps between cause and effect using

associations. Thus, association between cause and effect would consist of associations between the cause and first intermediate step, the first intermediate step and second intermediate step, and so on, while the overall association between cause and effect remain the same. This approach to mechanisms may be able to account for some effects of mechanism knowledge described earlier. For example, to account for why people believe more strongly in a causal link given a plausible mechanism for observed covariation (Fugelsang & Thompson, 2000), an advocate of associationism can argue that the mechanism conveys additional associative strength.

However, other effects of mechanism knowledge described earlier seem more challenging to the associationist approach. Ahn et al. (1995; Experiment 4) equated the covariation or association conveyed by the mechanism statements and the covariation statements, but participants nonetheless gave stronger causal attributions given the mechanism statements than covariation statements. Likewise, it is unclear on the associationist approach why conjunction and discounting effects are not obtained given purely covariational statements (Ahn & Bailenson, 1996) or why mechanism knowledge influences which categories we induce, given identical learning data (Hagmayer et al., 2011).

Forces and Powers

The associationist view contrasts most strongly with accounts of causal mechanisms in terms of *forces* (Talmy, 1988; Wolff, 2007) or *powers* (Harré & Madden, 1975; White, 1988, 1989). The intuition behind these approaches is that causal relations correspond to the operation of physical laws, acting on physical objects (Aristotle, 1970; Harré & Madden, 1975) or through physical processes (Dowe, 2000; Salmon, 1984; see also Danks, this volume). For example, Dowe (2000) argued that causal relations occur when a conserved quantity, such as energy, is transferred from one entity to another. This idea is broadly consistent with demonstrations that

people often identify visual collision events as causal or non-causal in ways concordant with the principles of Newtonian mechanics, such as conservation of momentum (Michotte, 1963/1946). Indeed, even young children seem to be sensitive to physical factors such as transmission in their causal reasoning (Bullock, Gelman, & Baillargeon, 1982; Shultz, Fisher, Pratt, & Rulf, 1986).

The *force dynamics* theory (Talmy, 1988; Wolff, 2007; Wolff, this volume) fleshes out these intuitions by representing causal relations as combinations of physical forces, modeled as vectors. On this theory, the causal *affector* (the entity causing the event) and *patient* (the entity operated on by the agent) are both associated with force vectors, indicating the direction of the physical or metaphorical forces in operation. For example, in a causal interaction between a fan and a toy boat, the fan would be the affector and the toy boat would be the patient, and both entities would have a vector indicating the direction of their motion. These forces as well as any other forces in the environment would combine to yield a *resultant* vector; e.g., the boat hits an orange buoy. On Wolff's (2007) theory, the affector *causes* a particular endstate to occur if (a) the patient initially does not have a tendency toward that endstate, but (b) the affector changes the patient's tendency, and (c) the endstate is achieved. For instance, the fan *caused* the boat to hit the buoy because (a) the boat was not initially headed in that direction, but (b) the fan changed the boat's course, so that (c) the boat hit the buoy. This sort of force analysis has been applied to several phenomena in causal reasoning, including semantic distinctions among causal vocabulary (*cause*, *enable*, *prevent*, *despite*; Wolff, 2007); the chaining of causal relations (e.g., *A* preventing *B* and *B* causing *C*; Barbey & Wolff, 2007); causation by omission (Wolff, Barbey, & Hausknecht, 2010); and direct versus indirect causation (Wolff, 2003).

A related physicalist approach is the *causal powers* theory (Harré & Madden, 1975; White, 1988, 1989). On this view, people conceptualize *particulars* (objects or persons) as

having dispositional causal properties, which operate under the appropriate *releasing conditions*. These properties can be either causal *powers* (capacities to bring about effects) or *liabilities* (capacities to undergo effects). For example, a hammer might strike a glass watch face, causing it to break (Einhorn & Hogarth, 1986). In this case, the hammer has a power to bring about breaking, and the glass has the liability to be broken. (See White, 2009b for a review of many studies consistent with the notion that causal relations involve transmission of properties among entities.) People then make causal predictions and inferences based on their knowledge of the causal powers and liabilities of familiar entities.

These physicalist theories capture a variety of intuitions and empirical results concerning causal thinking (see Waldmann & Mayrhofer, *in press*), and any complete theory of causal mechanisms is responsible for accounting for these phenomena. However, these theories are compatible with many different underlying representations. In the case of force dynamics, the vector representations are highly abstract and apply to *any* causal situation. That is, this theory does not posit representations for specific mechanisms in semantic memory, and therefore mechanism representations could take one of many formats. In the case of causal powers theory, the reasoner must represent properties of particular objects, which in combination could lead to representations of specific mechanisms. However, these property representations could potentially take several different representational formats, including icons and schemas (see below). Thus, although force and power theories certainly capture important aspects of causal reasoning, they do not provide a clear answer to the question of how mechanisms are mentally represented.

Icons

A related possibility is that people represent causal mechanisms in an iconic or image-like format. For example, when using mechanism knowledge to think about how a physical device works, the reasoner might mentally simulate the operation of the machine using mental imagery. More generally, people might store mechanism knowledge in an iconic format isomorphic to the physical system (Barsalou, 1999)—a view that sits comfortably with the physicalist theories described above. (Goldvarg & Johnson-Laird, 2001 propose a different, broadly iconic view of causal thinking based on mental models; see also Johnson-Laird & Khemlani, this volume.)

Forbus's (1984) *qualitative process theory* is an artificial intelligence theory of this style of reasoning. Qualitative process theory is designed to solve problems such as whether a bathtub will overflow, given the rate of water flowing out the faucet, the rate of drainage, and the rate of evaporation. This theory is 'qualitative' in the sense that it compares quantities and stores the direction of change, but does not reason about exact quantities. In this way, it is supposed to be similar to how humans solve these problems.

However, even if qualitative process theory accurately characterizes human problem solving processes, it is unclear whether these processes rely on mental representations that are propositional or image-like; after all, qualitative process theory itself is implemented in a computer programming language, using propositional representations. Several experimental results have been taken to support image-like representations (see Hegarty, 2004 for a review). First, when solving problems about physical causal systems (such as diagrams of pulleys or gears), participants who think aloud are likely to make gestures preceding their verbal descriptions, suggesting that spatial reasoning underlies their verbalizations (Schwartz & Black, 1996). Second, solving problems about physical causal systems (such as diagrams of pulleys or

gears) appears to rely on visual ability but not verbal ability. Performance on such problems is predicted by individual differences in spatial ability but not in verbal ability (Hegarty & Sims, 1994), and dual-task studies reveal interference between mechanical reasoning and maintenance of a visual working memory load but not a verbal working memory load (Sims & Hegarty, 1997).

It is an open question whether people run image-like mental simulations even when reasoning about causal processes that are less akin to physical systems, but some indirect support exists. For instance, asymmetries in cause-to-effect versus effect-to-cause reasoning suggest that people may use simulations. Tversky and Kahneman (1981) showed that people rate the conditional probability of a daughter having blue eyes given that her mother has blue eyes to be higher than the conditional probability of a mother having blue eyes given that the daughter has blue eyes. If the base rates of mothers and daughters having blue eyes are equal, these probabilities should be the same, but people appear to err because they make higher judgments when probability 'flows' with the direction of causality (for similar findings, see Fernbach, Darlow, & Sloman, 2010, 2011; Medin, Coley, Storms, & Hayes, 2003; Pennington & Hastie, 1988). While these results do not necessitate image-like representations, they do speak in favor of simulation processes, as forward simulations appear to be more easily 'run' than backward simulations, just as films with a conventional narrative structure are more readily understood than films like *Memento* in which the plot unfolds in reverse order.

However, other arguments and evidence suggest that these results may be better understood in terms of non-iconic representations. First, a number of researchers have argued that there are fundamental problems with iconic representations. Pylyshyn (1973) argues, for example, that if we store iconic representations and use them in the same way that we use visual

perception, then we need a separate representational system to *interpret* those icons, just as we do for vision. Rips (1984) criticizes mental simulation more generally, pointing out that the sort of mental simulation posited by AI systems in all but the simplest cases is likely to be beyond the cognitive capacity of human reasoners. Reasoning about turning gears is one thing, but Kahneman and Tversky (1982) claim that people use mental simulation to assess the probabilities of enormously complex causal systems, such as geopolitical conflict. Clearly, the number and variety of causal mechanisms at play for such simulations is beyond the ken of even the most sophisticated computer algorithms, much less human agents. In Rips's view, rule-based mechanisms are far more plausible candidates for physical causal reasoning. According to both Pylyshyn and Rips, then, the phenomenology of mental simulation may be epiphenomenal.

There is also empirical evidence at odds with iconic representations of mechanisms. For example, Hegarty (1992) gave participants diagrams of systems of pulleys, and asked them questions such as "if the rope is pulled, will pulley B turn clockwise or counterclockwise?" Response times were related to the number of components between the cause (here, the rope) and effect (pulley B). While this result is broadly consistent with the idea of mental simulation, it suggests that people simulate the system piecemeal rather than simultaneously (as one might expect for a mental image or 'movie'). More problematically, participants seem to be self-inconsistent when all parts are considered. In a study by Rips and Gentner (reported in Rips, 1984), participants were told about a closed room containing a pan of water. They were asked about the relations between different physical variables (such as air temperature, evaporation rate, and air pressure)—precisely the sort of inferences that mental simulations (such as those proposed by qualitative process theory) are supposed to be used for. The researchers found that people not only answered these questions inconsistently with the laws of physics, but even made

intransitive inferences. That is, participants very frequently claimed that a variable X causes a change in variable Y, which in turn causes a change in variable Z, but that X does not cause a change in Z—an intransitive inference. Such responses should not be possible if people are qualitatively simulating the physical mechanisms at work: Even if their mechanism knowledge diverges from the laws of physics, it should at least be internally consistent. (Johnson and Ahn (2015) review several cases where causal intransitivity can be normative, but none of these cases appear to be relevant to the stimuli used in the Rips and Gentner study). These results are more consistent with a schema view of mechanism knowledge (see below).

In sum, while studies of physical causal reasoning provide further evidence that causal thinking and mechanism knowledge in particular are used widely across tasks, they do not seem to legislate strongly in favor of iconic representations of mechanism knowledge. These results do, however, provide constraints on what representations could be used for mechanism-based reasoning.

Placeholders

A fourth representational candidate is a *placeholder* or *reference pointer*. On this view, people do not have elaborate knowledge about causal mechanisms underlying causal relations, but instead have a placeholder for a causal mechanism. That is, people would believe that every causal relation has an (unknown) causal mechanism, yet in most cases would not explicitly represent the content. (See Keil, 1989; Kripke, 1980; Medin & Ortony, 1989; Putnam, 1975 for the original ideas involving conceptual representations; and see Pearl, 2000 for a related, formal view.)

The strongest evidence for this position comes from *metacognitive illusions*, where people consistently overestimate their knowledge about causal systems (Rozenblit & Keil, 2002).

In a demonstration of the *illusion of explanatory depth* (IOED), participants were asked to rate their mechanistic knowledge of how a complex but familiar artifact operates (such as a flush toilet). Participants were then instructed to explain in detail how that artifact operates. When asked to re-rate their mechanistic knowledge afterwards, ratings were sharply lower, indicating that the act of explaining brought into awareness the illusory nature of their mechanistic knowledge. Thus, people's representations of causal mechanisms appear to differ from their metarepresentation—people's representations of mechanisms are highly skeletal and impoverished, yet their metarepresentations point to much fuller knowledge.

Further, this illusion goes beyond general overconfidence. Although similar effects can be found in other complex causal domains (e.g., natural phenomena such as how tides occur), people's knowledge is comparatively well-calibrated in non-causal domains, such as facts (e.g., the capital of England), procedures (e.g., how to bake chocolate chip cookies from scratch), and narratives (e.g., the plot of *Good Will Hunting*), although some (more modest) overconfidence can be found in these other domains too (Fischhoff, Slovic, & Lichtenstein, 1977).

Together, these results suggest that, at least in some cases, people do not store detailed representations of mechanisms in their heads, but rather some skeletal details together with a metarepresentational placeholder or 'pointer' to some unknown mechanism assumed to exist in the world. These impoverished representations, together with the robust illusions of their richness, are another reason to be suspicious of iconic representations of mechanism knowledge (see "Icons" above). To the extent that this is a plausible representational format because it feels introspectively right, we should be suspicious that this intuition may be a metacognitive illusion.

However, in addition to these metarepresentational pointers or placeholders, people clearly do have some skeletal representations of mechanisms. Many of the effects described in

earlier sections depend on people having some understanding of the content of the underlying mechanisms (e.g., Ahn & Bailenson, 1996; Ahn et al., 1995; Fugelsang & Thompson, 2000). And although people's mechanistic knowledge might be embarrassingly shallow for scientific phenomena and mechanical devices, it seems to be more complete for mundane phenomena. For instance, people often drink water after they exercise. Why? Because they become thirsty. Although the physiological details may elude most people, people surely understand this mechanism at a basic, skeletal level. If not as associations, causal powers, or icons, what format do these representations take? Below, we consider two possibilities for these skeletal representations—causal networks and schemas.

Networks

The idea that causal mechanisms might be represented as *networks* has recently received much attention (e.g., Glymour & Cheng, 1998; Griffiths & Tenenbaum, 2009; Pearl, 2000). According to this view, causal relationships are represented as links between variables in a directed graph, encoding the probabilistic relationships among the variables and the counterfactuals entailed by potential interventions (see Rottman, this volume for more details). For example, people know that exercising (*X*) causes a person to become thirsty (*Y*), which in turn causes a person to drink water (*Z*). The causal arrows expressed in the graph encode facts such as: (1) Exercising raises the probability that a person becomes thirsty (a probabilistic dependency); and (2) intervening to make a person exercise (or not exercise) will change the probability of thirst (a counterfactual dependency). The relationship between thirst (*Y*) and drinking water (*Z*) can be analyzed in a similar way. These two relationships can lead a reasoner to infer, transitively, a positive covariation between exercise (*X*) and drinking water (*Z*), and a counterfactual dependence between interventions on exercise and the probability of drinking

water (but see "Schemas" below for several normative reasons why causal chains can be intransitive). Similarly, the effects of drinking water will also have probabilistic and counterfactual relationships to exercise, as will the alternative causes of drinking water, and so on. These networks are used in artificial intelligence systems because they are economical and efficient ways of storing and reasoning about causal relationships (Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 1993).

If causal knowledge is represented in causal networks, then they could be reducible to the probabilistic dependencies and counterfactual entailments implied by the network. One proponent of this view is Pearl (1988), who argued that our knowledge is fundamentally about probabilities, and that causal relationships are merely shorthand for probabilistic relationships (though Pearl, 2000 argues for a different view; see "Open Questions" below). If causal relations are merely abbreviations of probabilistic relationships, we can define a mechanism for the causal relationship $X \rightarrow Z$ as a variable Y which, when conditioned on, makes the correlation between X and Z go to zero (Glymour & Cheng, 1998) so that the Markov condition is satisfied. That is, Y is a mechanism for $X \rightarrow Z$ if $P(Z|X) > P(Z|\sim X)$, but $P(Z|X,Y) = P(Z|\sim X,Y)$. The intuition here is the same as in mediation analysis in statistics—a variable Y is a full mechanism or mediator if it accounts for the entirety of the relationship between X and Z. As an example, Glymour and Cheng (p. 295) cite the following case (from Baumrind, 1983):

> The number of never-married persons in certain British villages is highly inversely correlated with the number of field mice in the surrounding meadows. [Marriage] was considered an established cause of field mice by the village elders until the mechanisms of transmission were finally surmised: Never-married persons bring with them a disproportionate number of cats.

In this case, the number of cats (Y) would be a mechanism that mediates the relationship between marriage (X) and field mice (Z) because there is no longer a relationship between marriage and

field mice when marriage is held constant. In the next section, we discuss limitations of conceptualizing mechanisms this way after describing the schema format.

Schemas

Finally, mechanism knowledge might be represented in the form of *schemas*—clusters of content-laden knowledge stored in long-term memory. Schemas are critical for inductive inference because they are general knowledge that can be used to instantiate many specific patterns (Bartlett, 1932; Schank & Abelson, 1977). For example, if Megan tells you about her ski trip, you can already fill in a great amount of the detail without her explicitly telling you—you can assume, for example, that there was a mountain, that the ground was snowy, that warm beverages were available in the lodge, and so on. Causal mechanisms could likewise be represented as clusters of knowledge about the underlying causal relations.

Like networks, schemas are a more skeletal representation and would not necessarily implicate image-like resources. Unlike networks, however, relationships between causally adjacent variables would not necessarily be stored together. This is because two causal relationships can be 'accidentally' united in a causal chain by sharing an event in common, yet not belong to the same schema. For example, we have a schema for sex causing pregnancy, and another schema for pregnancy causing nausea. But we may not have a schema for the relationship between sex and nausea. On the network view discussed above (Glymour & Cheng, 1998), because these three events are related in a causal chain, pregnancy is a mechanism connecting sex and nausea. On the schema view, in contrast, sex and nausea might not even be seen as causally related.

To distinguish between networks and schemas, Johnson and Ahn (2015) tested people's judgments about the *transitivity* of causal chains—the extent to which, given that *A* causes *B* and

B causes *C*, *A* is seen as a cause of *C*. According to the network view, the *A*→*C* relationship should be judged as highly causal to the extent that *A*→*B* and *B*→*C* are seen as highly causal. In contrast, the schema view implies that *A*→*C* would be judged as highly causal only if *A* and *C* belong to the same schema, even if *A*→*B* and *B*→*C* are strong. This is exactly what was found. For chains that were found in a preliminary experiment to be highly schematized (e.g., Carl studied, learned the material, and got a perfect score on the test), participants gave high causal ratings to *A*→*B*, *B*→*C*, and *A*→*C* (agreeing that Carl studying caused him to get a perfect score on the test). But for chains that were not schematized (e.g., Brad drank a glass of wine, fell asleep, and had a dream), participants gave high causal ratings for *A*→*B* and *B*→*C*, but not for *A*→*C* (denying that Brad's glass of wine made him dream). Johnson and Ahn (2015) also ruled out several normative explanations for causal intransitivity (e.g., Hitchcock, 2001; Paul & Hall, 2013). For example, causal chains can be normatively intransitive when the Markov condition is violated, but the Markov condition held for the intransitive chains. Similarly, chains can appear intransitive if one or both of the intermediate links (*A*→*B* or *B*→*C*) is probabilistically weak, because the overall relation (*A*→*C*) would then be very weak. But the transitive and intransitive chains were equated for intermediate link strength, so this explanation cannot be correct.

The lack of transitive inferences given unschematized causal chains is a natural consequence of the schema theory, but is difficult to square with the network theory. When assessing whether an event causes another, people often use a 'narrative' strategy, rejecting a causal relationship between two events if they cannot generate a story leading from the cause to the effect using their background knowledge (e.g., Kahneman & Tversky, 1982; Taleb, 2007). Hence, if people store *A*→*B* and *B*→*C* in separate schemas, they could not easily generate a path leading from *A* to *C*, resulting in intransitive judgments. The very point of the network

representation, however, is to allow people to make precisely such judgments—to represent, for example, the conditional independence between *A* and *C* given *B*, and the effects of potential interventions on *A* on downstream variables. Indeed, if the network view *defines* mechanisms in terms of such conditional independence relations, then it would require these variables to be linked together. Participants' intransitive judgments, then, are incompatible with network representations.

Open Questions

Because the issue of how causal knowledge is represented is a young research topic, we think it is fertile ground for further theoretical and empirical work. The greatest challenge appears to be understanding how mechanism knowledge can have all the representational properties that it does—it has schema-like properties (e.g., causally adjacent variables are not necessarily connected in a causal network; Johnson & Ahn, 2015), yet it also has association-like properties (e.g., causal reasoning sometimes violates probability theory in favor of associationist principles; Rehder, 2014), force-like properties (e.g., vector models capture aspects of causal reasoning; Wolff, 2007), icon-like properties (e.g., people have the phenomenology of visual simulation in solving mechanistic reasoning problems; Hegarty, 2004), placeholder-like properties (e.g., our metarepresentations are far richer than our representations of mechanisms; Rozenblit & Keil, 2002), and network-like properties (e.g., people are sometimes able to perform sophisticated probabilistic reasoning in accord with Bayesian networks; Gopnik et al., 2004).

One view is that Bayesian network theories will ultimately be able to encompass many of these representational properties (Danks, 2005). Although one version of the network theory equates mechanism knowledge with representing the causal graph (Glymour & Cheng, 1998),

other network-based theories might be more flexible (e.g., Griffiths & Tenenbaum, 2009). For example, Pearl (2000, p. xv–xvi) writes:

> In this tradition [of Pearl's earlier book *Probabilistic Reasoning in Intelligent Systems* (1988)], probabilistic relationships constitute the foundations of human knowledge, whereas causality simply provides useful ways of abbreviating and organizing intricate patterns of probabilistic relationships. Today, my view is quite different. I now take causal relationships to be the fundamental building blocks both of physical reality and of human understanding of that reality, and I regard probabilistic relationships as but the surface phenomena of the causal machinery that underlies and propels our understanding of the world.

That is, our causal knowledge might be represented on two levels—at the level of causal graphs that represent probabilities and counterfactual entailments, and at a lower level that represents the operation of physical causal mechanisms. This view does not seem to capture all of the empirical evidence, as the results of Johnson and Ahn (2015) appear to challenge any theory that posits representations of causal networks without significant qualifications. Nonetheless, theories that combine multiple representational formats and explain the relations among them are needed to account for the diverse properties of mechanism knowledge.

Another largely open question is where the content of these representations comes from. For example, to the extent that mechanism knowledge is stored in a schema format, where do those schemas come from? That is, which event categories become clustered together in memory and which do not? Little is known about this, perhaps because schema formation is multiply determined, likely depending on factors such as spatial and temporal contiguity, frequency of encounter, and others. This problem is similar in spirit and difficulty to the problem of why we have the particular concepts that we do. Why do we have the concept of "emerald" but not the concept of "emeruby" (an emerald before 1997 or a ruby after 1997; Goodman, 1955)? Likewise, why do we have a schema for pregnancy and a schema for nausea, but not a schema that

combines the two? Although we describe prior research below on how people learn causal mechanisms, this existing work does not resolve the issue of where causal schemas come from.

Learning Causal Mechanisms

In this section, we address how mechanism knowledge is learned. Associationist and network theories have usually emphasized learning from statistical induction (e.g., Glymour & Cheng, 1998). However, these theories can also accommodate the possibility that much or even most causal knowledge comes only *indirectly* from statistical induction. For example, some mechanisms could have been induced by our ancestors and passed to us by cultural evolution (and transmitted by testimony and education) or biological evolution (and transmitted by the selective advantage of our more causally enlightened ancestors). Although the bulk of empirical work on the acquisition of mechanisms focused on statistical induction, we also summarize what is known about three potential indirect learning mechanisms—testimony, reasoning, and perception.

Direct Statistical Induction

If mechanisms are essentially patterns of covariation, as some theorists argue (Glymour & Cheng, 1998; Pearl, 1988), then the most direct way to learn about mechanisms is by inducing these patterns through statistical evidence. In fact, people are often able to estimate the probability of a causal relationship between two variables from contingency data (e.g., Griffiths & Tenenbaum, 2005; see also Rottman, this volume). However, mechanisms involve more than two variables, and the ability to learn causal relationships from contingency data largely vanishes when additional variables are introduced. For instance, in Steyvers, Wagenmakers, Blum, and Tenenbaum (2003), participants were trained to distinguish between three-variable common cause (i.e., *A* causes both *B* and *C*) and common effect (i.e., *A* and *B* both cause *C*). Although

performance was better than chance levels (50% accuracy), it was nonetheless quite poor—less than 70% accuracy on average even after 160 trials, with nearly half of participants performing no better than chance. (For similar results, see Hashem & Cooper, 1998 and White, 2006.) Although people are better able to learn from intervention than from mere observation (Kushnir & Gopnik, 2005; Lagnado & Sloman, 2004; Waldmann & Hagmayer, 2005; see also Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015), they are still quite poor at learning multivariable causal structures. In Steyvers et al. (2003), learners allowed to intervene achieved only 33% accuracy at distinguishing among the 18 possible configurations of three variables (compared to 5.6% chance performance and 100% optimal performance). For the complex causal patterns at play in the real world, it seems unlikely that people rely on observational or interventional learning of multivariable networks as their primary strategy for acquiring mechanism knowledge.

Given that people have great difficulty learning a network of only 3 variables when presented simultaneously, a second potential learning strategy is *piecemeal* learning of causal networks. That is, instead of learning relations among multiple variables at once, people may first acquire causal relationships between two variables, and then combine them into larger networks (Ahn & Dennis, 2000; Fernbach & Sloman, 2009). For example, Baetu and Baker (2009) found that people who learned a contingency between *A* and *B* and between *B* and *C* inferred an appropriate contingency between *A* and *C*, suggesting that participants had used the principle of causal transitivity to combine inferences about these disparate links (for similar findings, see Goldvarg & Johnson-Laird, 2001; von Sydow, Meder, & Hagmayer, 2009).[1]

[1] Although this result may appear to conflict with the results of Johnson and Ahn (2015), which demonstrated causal intransitivity in some causal chains, the two sets of findings can be reconciled, because Johnson and Ahn (2015) used familiar stimuli for which people could expect

Although more work will be necessary to test the boundary conditions on piecemeal construction of causal networks (e.g., Johnson & Ahn, 2015), this appears to be a more promising strategy for acquiring knowledge of complex causal mechanisms.

Learning networks of causal relations from contingency data is challenging, whether from observations or from interventions, likely as a result of our computational limits. Hence, it seems unlikely that we induce all of our mechanism knowledge from statistical learning (see Ahn & Kalish, 2000), even if direct statistical induction plays some role. Where might these other beliefs about causal mechanisms come from?

Indirect Sources of Mechanism Knowledge

Much of our mechanism knowledge appears to come not directly from induction over observations, but from other sources, such as testimony from other people or explicit education, reasoning from other beliefs, and perhaps perception. Although relatively little work has addressed the roles of these sources in acquiring *mechanism* knowledge in particular, each has been implicated in causal learning more generally.

Testimony and cultural evolution. Much of our mechanism knowledge seems to come from family members and peers, from experts, and from formal and informal education. Children are famously curious, and renowned for their enthusiasm for asking series of "why" questions that probe for underlying mechanisms. Although parents are an important resource in children's learning (e.g., Callanan & Oakes, 1992), parents' knowledge is necessarily limited by their expertise. However, children's (and adults') ability to seek out and learn from experts puts them

---

to have schematized knowledge, whereas Baetu and Baker (2009) used novel stimuli. In reasoning about novel stimuli, people would not use a narrative strategy (i.e., trying to think of a story connecting the causal events), but would instead use a statistical (Baetu & Baker, 2009) or rule-based strategy (Goldvarg & Johnson-Laird, 2001). The lack of schematized knowledge would not block transitive inferences under these reasoning strategies.

in a position to acquire mechanism knowledge when unavailable from more immediate informants (Mills, 2013; Sobel & Kushnir, 2013; Sperber et al., 2010). In particular, children have an understanding of how knowledge is distributed across experts (Lutz & Keil, 2002) and which causal systems are sufficiently rich or "causally dense" that they would have experts (Keil, 2010).

Further, the growth of mechanism knowledge not only over ontogeny but over history points to powerful mechanisms of cultural evolution (Boyd & Richerson, 1985; Dawkins, 1976). Successive generations generate new scientific knowledge and transmit a subset of that knowledge to the public and to other scientists. Most experimental and computational work in cultural evolution has focused on how messages are shaped over subsequent generations (Bartlett, 1932; Griffiths, Kalish, & Lewandowsky, 2008), how languages evolve (Nowak, Komarova, & Niyogi, 2001), or how beliefs and rituals are propagated (Boyer, 2001). Less is known from a formal or experimental perspective about how cultural evolution impacts the adoption of scientific ideas (but see Kuhn, 1962). Nonetheless, it is clear that the succession of ideas over human history are guided in large part by a combination of scientific scrutiny and cultural selection, and that these forces therefore contribute to the mechanism knowledge that individual cognizers bring to bear on the world.

Reasoning. Imagine you have done the hard work of understanding the mechanisms underlying the circulatory system of elephants—perhaps by conducting observations and experiments, or through explicit education. It would be sad indeed if this hard-won mechanism knowledge were restricted to causal reasoning about elephants. What about specific kinds of elephants? Mammals in general? Particular mammals like zebras?

Beliefs are not informational islands. Rather, we can use reasoning to extend knowledge from one domain to another. We can use *deductive* reasoning to extend our general knowledge about elephant circulation 'forwards' to African elephant circulation (Johnson-Laird & Byrne, 1991; Rips, 1994; Stenning & van Lambalgen, 2008; see Oaksford & Chater, this volume and Over, this volume). We can use *analogical* reasoning to extend our knowledge of elephant circulation 'sideways' to similar organisms like zebras (Gentner & Markman, 1997; Hofstadter, 2014; Holyoak & Thagard, 1997; see Holyoak & Lee, this volume). And we can use *abductive* reasoning to extend our knowledge 'backwards' to mammals (Keil, 2006; Lipton, 2004; Lombrozo, 2012; see Lombrozo & Vasilyeva, this volume and Meder & Mayrhofer, this volume); indeed, Ahn and Kalish (2000) suggested that abductive reasoning is a particularly important process underlying mechanistic causal reasoning. Although these reasoning strategies do not always lead to veridical beliefs (e.g., Lipton, 2004; Stenning & van Lambalgen, 2008), they seem to do well often enough that they can be productive sources of hypotheses about causal mechanisms, and they may be accurate enough to support causal inference in many realistic circumstances without exceeding our cognitive limits.

Perception. Intuitively, we sometimes seem to learn mechanisms from simply watching those mechanisms operate in the world (see White, this volume). For example, you might observe a bicycle in operation, and draw conclusions about the underlying mechanisms from these direct observations. Indeed, much evidence supports the possibility that people can visually perceive individual causal relations (Michotte, 1963/1946; Rolfs, Dambacher, & Cavanagh, 2013; see White, 2009a for a review and Rips, 2011 for a contrary view). Haptic experiences may also play a role in identifying causal relations (White, 2012, 2014; Wolff & Shepard, 2013). Just as people seem to learn about individual causal relationships from statistical information and

combine them together into more detailed mechanism representations (Ahn & Dennis, 2000; Fernbach & Sloman, 2009), people may likewise be able to learn about individual causal events from visual experience, and combine these into larger mechanism representations.

However, we should be cautious in assuming that we rely strongly on perceptual learning for acquiring mechanism knowledge, because little work has addressed this question directly and people are susceptible to metacognitive illusions (Rozenblit & Keil, 2002). For example, Lawson (2006) found that people have poor understanding of how bicycles work, and when asked to depict a bicycle from memory, often draw structures that would be impossible to operate (e.g., because the frame would prevent the wheels from turning). These errors were found even for bicycle experts and people with a physical bicycle in front of them while completing the task (see also Rozenblit & Keil, 2002). Hence, in many cases, what appears to be a mechanism understood through direct perceptual means is in fact something far more schematic and incomplete, derived from long-term memory.

Open Questions

One major open question concerns the balance among these direct and indirect sources. Do we acquire many of our mechanism beliefs through statistical induction, despite our difficulty with learning networks of variables, or is the majority of our causal knowledge derived from other indirect sources? When we combine individual causal relations into mechanism representations, do we do so only with relations learned statistically, or are we also able to combine disparate relations learned through testimony, reasoning, or perception? To what extent can these causal maps combine relations learned through *different* strategies? Put differently, do these learning strategies all produce mechanism representations of the same format, or do they contribute different sorts of representations that may be difficult to combine into a larger picture?

Another challenge for future research will be investigating the extent to which these sources contribute not only to learning general causal knowledge (learning that *A* causes *B*) but also mechanism knowledge (learning *why A* causes *B*). The majority of the evidence summarized above concerns only general causal knowledge, so the contribution of these indirect sources to acquiring mechanism knowledge should be addressed empirically.

Finally, might some mechanism knowledge be conveyed through the generations not only through cultural evolution, but also through biological evolution? It is controversial to what extent we have innate knowledge (e.g., Carey, 2009; Elman et al., 1996), and less clear still to what extent we have innate knowledge of causal mechanisms. Nonetheless, we may be born with some highly schematic, skeletal representations of mechanisms. For example, 4-month-old infants appear to understand the fundamental explanatory principles of physics (e.g., Spelke, Breinlinger, Macomber, & Jacobson, 1992), including physical causality (Leslie & Keeble, 1987); belief-desire psychology emerges in a schematic form by 12 months (Gergely & Csibra, 2003); and young children use the principles of essentialism (Keil, 1989), vitalism (Inagaki & Hatano, 2004), and inherence (Cimpian & Salomon, 2014) to understand the behavior of living things. These rudimentary explanatory patterns may provide candidate mechanisms underlying many more specific causal relationships observed in the world. To the extent that these patterns are innate, we might be born with some highly skeletal understanding of causal mechanisms that can underlie later learning.

Conclusion

The chapters in this volume demonstrate the depth to which causality pervades our thinking. In this chapter, we have argued further that knowledge of causal mechanisms pervades our causal understanding. First, when deciding whether a relationship is causal, mechanism

knowledge can trump other cues to causality. It provides evidence over and above covariation, and a mechanism can even change the interpretation of new covariation information; it can result in violations of the Causal Markov Condition—a critical assumption for statistical reasoning via Bayesian networks; and it can alter expectations about temporal delays, moderating the effect of temporal proximity on causal judgment. Second, mechanism knowledge is crucial to inductive inference. It affects which categories are used and induced; how strongly an exemplar's features are projected onto other exemplars; how likely we are to extend a property from one category to another; and how we make category-based probability judgments, producing discounting and conjunction effects.

Mechanism knowledge is also key to how causal relations are mentally represented. Several representational formats have been proposed—associations, forces or powers, icons, placeholders, networks, and schemas. Although there are likely to be elements of all of these formats in our mechanism knowledge, two positive empirical conclusions are clear: First, people's metarepresentations of causal knowledge are far richer than their actual causal knowledge, suggesting that our representations include abstract placeholders or 'pointers' to real-world referents that are not stored in the head. Second, however, people *do* represent some mechanism content, and this content appears to often take the form of causal schemas. Future theoretical and empirical work should address how the various properties of mechanism knowledge can be understood in a single framework.

Mechanisms may be acquired in part through statistical induction. However, because people are poor at learning networks of three or more variables by induction, it is more likely that people learn causal relations individually and assemble them piecemeal into larger networks. People also seem to use other learning strategies for acquiring mechanism knowledge, such as

testimony, reasoning, and perhaps perception. How these strategies interact, and whether they produce different sorts of representations, are open questions.

Although we would not claim that all reasoning about causation is reasoning about mechanisms, mechanisms are central to many of our nearest and dearest inferential processes. Hence, understanding the representation and acquisition of mechanism knowledge can help to cut to the core of causal thinking, and much of the cognition that it makes possible.

References

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*, 82–123.

Ahn, W., & Dennis, M. J. (2000). Induction of causal chains. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 19–24). Mahwah, NJ: Erlbaum.

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 199–226). Cambridge, MA: MIT Press.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Aristotle (1970). *Physics, Books I–II.* W. (Trans. W. Charlton.) Oxford, UK: Clarendon Press.

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*, 153–168.

Barbey, A. K., & Wolff, P. (2007). Learning causal structure from reasoning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 713–718). Austin, TX: Cognitive Science Society.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge, UK: Cambridge University Press.

Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology*, *45*, 1289–1298.

Boyd, R., & Richerson, P. J. (2005). *The origin and evolution of cultures*. Oxford, UK: Oxford University Press.

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York, NY: Basic Books.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 708–731.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning*, *8*, 269–293.

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology*, *56A*, 865–890.

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *The Quarterly Journal of Experimental Psychology*, *57B*, 179–191.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*, 353–378.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York, NY: Academic Press.

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*, 213–233.

Carey, S. (2009). *The origin of concepts*. Oxford, UK: Oxford University Press.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*, 461–527.

Coenen, A., Rehder, B., Gureckis, T. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology, 79,* 102–133.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Danks, D. (2005). The supposed competition between theories of human causal inference. *Philosophical Psychology*, *18*, 259–272.

Dawkins, R. (1976). *The selfish gene*. Oxford, UK: Oxford University Press.

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

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*, 329–36.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*, 168–85.

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552–564.

Focht III, D. R., Spicer, C., & Fairchok, M. P. (2002). The efficacy of duct tape vs cryotherapy in the treatment of verruca vulgaris (the common wart). *Archives of Pediatrics & Adolescent Medicine*, *156*, 971–974.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.

Fugelsang, J. A., & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology*, *54*, 15–32.

Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology*, *58*, 86–95.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, *7*, 287–292.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*, 49–71.

Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probability: A normative approach. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 295–313). Oxford, UK: Oxford University Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 3–32.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Proceedings of the Royal Society, B*, *363*, 3503–3514.

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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Hagmayer, Y., Meder, B., von Sydow, M., & Waldmann, M. R. (2011). Category transfer in sequential causal learning: The unbroken mechanism hypothesis. *Cognitive Science*, *35*, 842–873.

Harré, R., & Madden, E. H. (1975). *Causal powers: A theory of natural necessity*. Lanham, MD: Rowman & Littlefield.

Hashem, A. I., & Cooper, G. F. (1996). Human causal discovery from observational data. *Proceedings of the AMIA Annual Fall Symposium*, 27–31.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1084–1102.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*, 280–285.

Hegarty, M., & Sims, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory & Cognition*, *22*, 411–430.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*, 273–299.

Hofstadter, D. R. (2014). *Surfaces and essences: Analogy as the fuel and fire of thought*. New York, NY: Basic Books.

Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist*, *52*, 35–44.

Hume, D. (1977). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett. (Original work published 1748.)

Inagaki, K., & Hatano, G. (2004). Vitalistic causality in young children's naive biology. *Trends in Cognitive Sciences*, *8*, 356–362.

Johnson, S. G. B., & Ahn, W. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*.

Johnson, S. G. B., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, *143*, 2223–2241.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction: Essays in cognitive psychology*. Hillsdale, NJ: Erlbaum.

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

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254.

Keil, F. C. (2010). The feasibility of folk science. *Cognitive Science*, *34*, 826–862.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107–128.

Kemp, C., & Jern, A. (2013). A taxonomy of inductive problems. *Psychonomic Bulletin & Review*, *21*, 23–46.

Kowlowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kripke, S. (1980). *Naming and necessity*. Oxford, UK: Blackwell.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*, 678–683.

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

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

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, *34*, 1667–1675.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*, 265–288.

Lipton, P. (2004). *Inference to the best explanation* (2nd Edition). London, UK: Routledge.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–32.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford, UK: Oxford University Press.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*, 167–204.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the decision of cognitive labor. *Child Development*, *73*, 1073–1084.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25.

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.

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

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, UK: Cambridge University Press.

Michotte, A. (1963). *The perception of causality*. (T. R. Miles & E. Miles, Trans.). New York, NY: Basic Books. (Original work published 1946.)

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology*, *49*, 404–418.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 904–919.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In *Advances in cognitive science, vol. 2: Theory and applications* (pp. 23–45). Chichester, UK: Ellis Horwood.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*, 114–118.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*, 186–216.

Park, J., & Sloman, S. A. (2014). Causal explanation in the face of contradiction. *Memory & Cognition*, *42*, 806–820.

Patalano, A. L., & Ross, B. H. (2007). The role of category coherence in experience-based prediction. *Psychonomic Bulletin & Review*, *14*, 629–634.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford, UK: Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

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

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 521-533.

Putnam, H. (1975). The meaning of "meaning." In K. Gunderson (Ed.), *Language, mind, and knowledge* (pp. 131–193). Minneapolis, MN: University of Minnesota Press.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*, 1–24.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107.

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

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*, 113–153.

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1261–1275.

Rips, L. J. (1984). Mental muddles. In M. Brand & R. M. Harnish (Eds.), *The representation of knowledge and belief*. Tuscon, AZ: University of Arizona Press.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Rips, L. J. (2011). Causation from perception. *Perspectives on Psychological Science*, *6*, 77–97.

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, *23*, 250–254.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*, 109–139.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*, 93–125.

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

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

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. New York, NY: Psychology Press.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, *35*, 303–317.

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science*, *20*, 457–497.

Shanks, D. R. (1987). Associative accounts of causality judgment. *Psychology of Learning and Motivation*, *21*, 229–261.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*, 139–159.

Shultz, T. R., Fisher, G. W., Pratt, C. C., & Rulf, S. (1986). Selection of causal rules. *Child Development*, *57*, 143–152.

Simon, H. A. (1996). *The sciences of the artificial* (3rd Edition). Cambridge, MA: MIT Press.

Sims, V. K., & Hegarty, M. (1997). Mental animation in the visuospatial sketchpad: Evidence from dual-task studies. *Memory & Cognition*, *25*, 321–332.

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

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*, 779–797.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*, 605–632.

Spellman, B. A., Price, C. M., & Logan, J. M. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition*, *29*, 193–208.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*, 359–393.

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

Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.

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

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York, NY: Random House.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*, 49–100.

Tversky, A., & Kahneman, D. (1981). *Evidential impact of base rates*. Technical Report. Office of Naval Research.

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

Von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In N. A. Taatgen & H. van Rign (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 803–808). Austin, TX: Cognitive Science Society.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 53–76.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 216–227.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*, 27–58.

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

Waldmann, M. R., & Mayrhofer, R. (*in press*). Hybrid causal representations. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 65). San Diego, CA: Academic Press.

White, P. A. (1988). Causal processing: Origins and development. *Psychological Bulletin*, *104*, 36–52.

White, P. A. (1989). A theory of causal processing. *British Journal of Psychology*, *80*, 431–454.

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

White, P. A. (2009a). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*, 580–601.

White, P. A. (2009b). Property transmission: An explanatory account of the role of similarity information in causal inference. *Psychological Bulletin*, *135*, 774–793.

White, P. A. (2012). The experience of force: The role of haptic experience of forces in visual perception of object motion and interactions, mental simulation, and motion-related judgments. *Psychological Bulletin*, *138*, 589–615.

White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science*, *38*, 38–75.

Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*, 1–48.

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

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*, 191–221.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. *Psychology of Learning and Motivation*, *58*, 167–202.