Précis of *Cognition as Sense-Making*

Samuel G.B. Johnson

Chapter One: Of the Sense-Making Faculties

Curiosity may have killed the cat, but it is also what keeps us humans alive. The facts most critical to survival are often hidden beneath the surface. Why did that driver swerve?—is there an obstacle on the road ahead? Why did that salesperson copy my credit card number?—is she swindling me? What made that creaking sound upstairs—is it an intruder, a ghost, or the wind?

Hidden explanations often entail courses of action, and humans must have strategies for arriving at these explanations and translating them into predictions and choices. Explanatory (or abductive) reasoning was first carefully characterized by the philosopher C.S. Peirce (1903). In explanatory inferences, we think of a premise that would explain an observation and infer that premise as true. This can be contrasted to *deductive reasoning*—where the premises entail the conclusion with certainty—in that the conclusions of explanatory inference are not logically entailed. It can also be contrasted to *inductive reasoning*—where repeated regularities are generalized—in that explanatory inferences can be made from a single observation.

Other species of animal probably do not devote the sort of energy that we do to making sense of their experiences. Even primates have only limited abilities to reason about cause and effect (Premack & Premack, 1994) and they fail many theory-of-mind tasks (Call & Tomasello, 2008). Given that other animals seem to get along rather well without such inferences, you might be forgiven for treating sense-making as epiphenomenal—a cognitive luxury that makes for better dinner conversation and a livelier time at the movies, but contributes little to "serious" cognition.

I believe that this view is mistaken—that our sense-making capacity is an indispensable source of beliefs about the world, about each other, and about ourselves; that, as a social species and as a problem-solving species, this capacity is fundamental to our survival; and that without this capacity, our mental lives would bear little resemblance to human cognition. More than any other type of thought pattern, I believe that sense-making is at the foundation of what makes us "rational animals."

The idea that explanation is psychologically important is not new. Prominent researchers across many sub-disciplines of cognitive science have posited that this-or-that capacity is "explanatory"—such claims have been made, in various guises, not only for high-level cognitive processes such as categorization and causal inference, but for such diverse faculties as stereotyping, mental-state inference, emotion, self-knowledge, perception, language, and memory. What *is* new is the idea that there is something in common to these explanatory processes—not merely a logic of perception, a logic of memory, a logic of categorization, a logic of mental-state inference, but a *logic of explanation*, which has domain-general inputs into many of these capacities. Deductive and inductive reasoning have each been the subject of multiple, incommensurable efforts to identify an underlying general logic, along with accompanying

empirical litigation (e.g., Holland et al., 1989; Johnson-Laird & Byrne, 1991; Rips, 1994; Sloman, 1993; Tenenbaum, Griffiths, & Kemp, 2006). Yet, far less is known about how explanatory logic works, or indeed *could* work. Humans solve explanatory problems that baffle not only other animals but also supercomputers—problems so baffling that we do not even know how to study them.

My objective is twofold: To elaborate some of the mechanisms that underlie explanatory processes, and to show that these same mechanisms, and their accompanying biases, are used across several such faculties—that these faculties share an underlying *explanatory logic*. That is, the mind faces similar problems in understanding language, inferring the causes of events, assessing the reasons for others' behavior, and so on. Although it is possible that each of these processes accomplishes explanatory inference in a different way, it is at least plausible that the mind has developed the same solutions to these common computational problems.

But where might one begin to look for common principles of explanatory logic? My approach is to consider puzzles the mind must solve in explanatory reasoning, and to evaluate their possible solutions as hypotheses in empirical psychology.

The first constraint is that explanatory inferences must be reasonably *veridical*. Cognition is useful to an organism because it produces a more-or-less accurate representation of the external world. In line with a long tradition in psychology and philosophy, I adopt a probabilistic standard of truth. That is, I assume that, normatively, beliefs reflect subjective probabilities. The mathematics of Bayesian probability—updating prior beliefs in light of their fit to new evidence—provides an elegant way to express normative explanatory inferences.

However, Bayesian rationality poses a number of challenges for limited minds in a limited environment. We face a problem of *informational poverty* because we face incomplete evidence for solving explanatory problems. We face a problem of *indeterminacy* because we often lack a principled way of assigning priors and likelihoods. And we face problems of *computational limits* because one inference often depends on another, leading to computational explosions. Chapters 2–4 examine strategies people use for solving these problems, while Chapters 5–7 look at the developmental origins and generality of these strategies across cognition.

Chapter Two: Of the Unseen

We must often make sense of things in the face of incomplete evidence. For example, doctors diagnose diseases when some test results are unavailable, and juries assess the credibility of a defendant's story based on sketchy evidence and conflicting testimony. And every day, we all infer others' mental states based on scant clues, infer the categories of objects based on occluded views, and infer causes when many of their potential effects are unknown.

Chapter 2 argues that people solve this *informational poverty* problem in part by relying on both observed evidence and *inferred evidence* about potentially diagnostic cues. For example, suppose a doctor must assess whether a patient has one of two equally common diseases (see Figure 1), given that the patient has one known symptom (symptom X) but another symptom is

unknown (symptom Z). A disease that causes both symptoms (disease H_W) would have *wide latent scope* because it makes unverified predictions, whereas a disease that causes only symptom X (H_N) would have *narrow latent scope* because its predictions are all verified. Normatively, the diseases are equally likely because they have the same base rates and no evidence distinguishes between them. Might people nonetheless make *inferences* about Z, leading them to favor one hypothesis over the other?

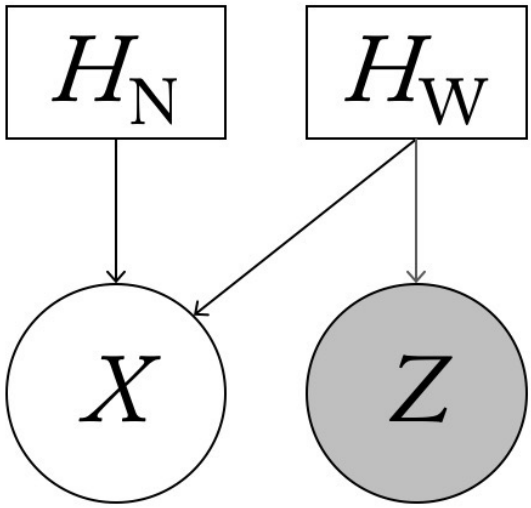

Figure 1. Causal structure where diagnostic evidence (Z) is unavailable.

Several pieces of converging evidence indicate an affirmative answer. First, people rely on the base rates of the unknown evidence, P(Z), to attempt to *guess* whether Z would be observed in the current case. P(Z) is irrelevant if the base rates of the causes, P(H_W) and P(H_N) are known, since a patient having either H_W or H_N has exactly a 50% chance of having Z unless there are alternative causes of Z (which are irrelevant to distinguishing between H_W and H_N). Nonetheless, Studies 1 and 2 found that whereas people favor H_N when P(Z) is low (say, 5%), as found in prior studies (Khemlani et al., 2011), they actually favor H_W when P(Z) is high. Thus, people erroneously use P(Z) to infer whether Z is present in the current case. Studies 2 and 3 found that this effect of evidence base rates cannot be explained by normative inferences about priors or by conversational pragmatics. Further, Study 4 found that people are motivated to seek information about unknown effect base rates, but not about known effect base rates. This stands in contrast to the laws of probability, according to which neither base rate is diagnostic if the *cause* base rates are known.

Similar effects were found in categorization. Participants in Study 5 read about individuals with a known feature (e.g., white spots) and an unknown feature (e.g., hollow antlers) and were asked to classify them into wide categories that had the unknown feature (e.g., *trocosiens* deer) or

narrow categories (*myronisus* deer) that lacked the feature. The base rate of the unknown feature was manipulated by indicating that the trait was either common or uncommon among *other* (irrelevant) species of deer. Once again, participants favored the narrow category to a greater degree when the unknown feature had a low base rate. Thus, the inferred evidence strategy appears to be widespread in cognition, appearing in the superficially distinct, but both fundamentally explanatory, processes of causal explanation and categorization.

Although this use of inferred evidence leads to bias, this is a side effect of a crucial and adaptive ability to identify diagnostic evidence and to selectively bring background knowledge to bear on it. Although people overextend this ordinarily useful strategy to irrelevant cues such as evidence base rates, people *do* often have knowledge that is useful for inferring diagnostic evidence (e.g., known features that are correlated with an unknown feature).

Beyond causal reasoning and categorization, inferred evidence is commonplace in perception, where it is used to infer contours (Kanizsa, 1976) and continuities of objects (Michotte, Thinès, & Crabbé, 1964), and generally to infer the three-dimensional world from a two-dimensional retinal array (Marr, 1982). This strategy may thus be quite general across explanatory processes and a crucial tool for solving the informational poverty problem.

Chapter Three: Of Simplicity and Complexity

Rational inferences require us to estimate an explanation's prior probability and its fit to the data. Unfortunately, we often lack a principled way to estimate these quantities. How can we even enumerate, much less assign probabilities to, the potential causes of a friend's frown, a movement in financial markets, or the fall of a civilization? Yet, we must have strategies for solving this *indeterminacy* problem, since we happily make inferences about all of these events.

One useful strategy is a *simplicity heuristic*—assuming that simpler explanations are generally better (Lombrozo, 2007). A doctor diagnosing a patient would favor a one-disease explanation (disease *A*) over a multiple-disease explanation (diseases *B* and *C*; see Figure 2). This preference is adaptive, because simpler explanations generally have higher prior probability. Yet, an explanation's posterior depends on both its prior and its fit to the data, and complex explanations often fit more tightly: Patients may be likelier to exhibit symptoms if they have two diseases rather than one. Thus, the optimal simplicity of an explanation follows a U-shaped curve: Too simple, and it does not account for the data; too complex, and it has a low prior and overfits (Forster & Sober, 1994).

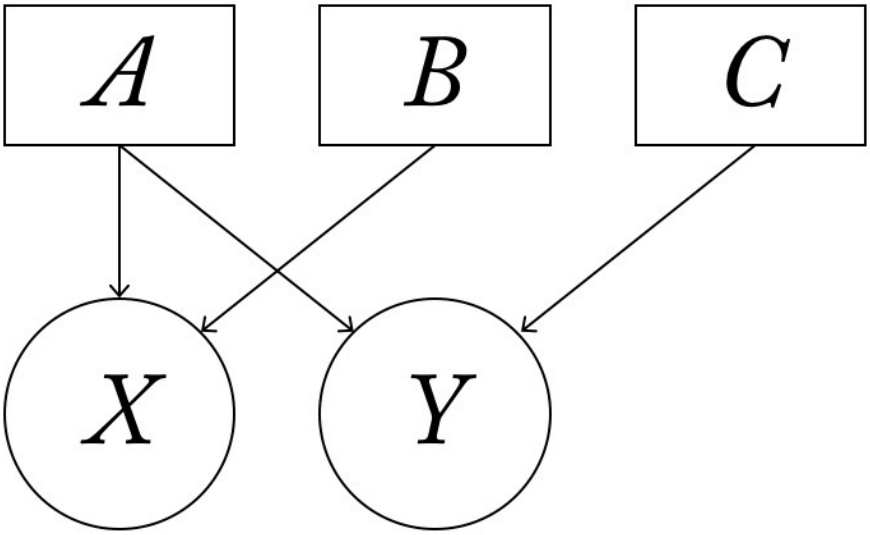

Figure 2. Causal structure with simple (*A*) and complex (*B* and *C*) explanations.

Chapter 3 argues that people manage this trade-off by using an *opponent heuristic* strategy. Most directly, participants in Study 9 indicated explicitly that simpler explanations had higher priors, P(*H*), but that *complex* explanations had higher likelihoods, P(*E*|*H*). These opponent heuristics allow people to use an easily available cue (simplicity) to simultaneously assess both of the often-indeterminate ingredients in Bayesian hypothesis comparison.

However, this raises a corollary problem: How can such a strategy yield inferences definite enough to provide a unique answer for a given problem, but flexible enough to provide different answers to different problems? The opponent heuristic strategy solves this dilemma by modulating the inference across contexts. Studies 10 and 11 found that people shift toward complex explanations in stochastic contexts, where likelihoods are imperfect, rendering a complexity heuristic computationally relevant. Studies 12 and 13 found that people favor simple explanations to varying degrees across domains, tracking general expectations about each domain's causal texture: Physical, rather than social, causal systems are thought to have more linear patterns of causality, leading to a stronger preference for simple explanations.

These opponent heuristics are used not only in obviously high-level tasks such as causal inference, but also in seemingly lower-level visual tasks such as assessing the fit of curves to scatterplot data. Whereas people should choose curves that are too simple (e.g., a linear curve for data generated by a quadratic function) if they use only a simplicity heuristic, Study 14 actually found that people choose curves that are too *complex* (e.g., a cubic curve). Further, Study 16 found that this bias occurs due to an *illusion of fit* wherein more complex curves are seen as better fits to the data, even if the simple and complex curves are equally close fits.

Causal explanation and visual curve-fitting, though superficially rather different, share a similar logical structure, and the mind relies on a similar heuristic logic for solving both

problems. Because we often lack principled ways to assign probabilities, such a strategy may be adaptive even as it leads to cognitive (and perhaps even perceptual) illusions. In novel situations where we cannot calculate but must simply guess, the best we can do is to adopt rules that work reasonably well most of the time.

Chapter Four: Of Probability and Belief

Our beliefs often entail other beliefs. For instance, knowing the thoughts of a central banker helps us to predict the future behavior of financial markets. However, such beliefs are often accompanied by uncertainty—perhaps we believe the central banker has a 70% chance of a positive outlook and a 30% chance of a negative one. Representing and using these degrees of belief is the very purpose of probability theory. Unfortunately, uncertainty propagates along a chain of reasoning. If the central bank's monetary policy choice depends probabilistically on the thoughts of the central banker, and congress's fiscal policy choice depends probabilistically on the bank's monetary policy choice, ad infinitum, we soon encounter a computational explosion in estimating the probability of events down the inference chain.

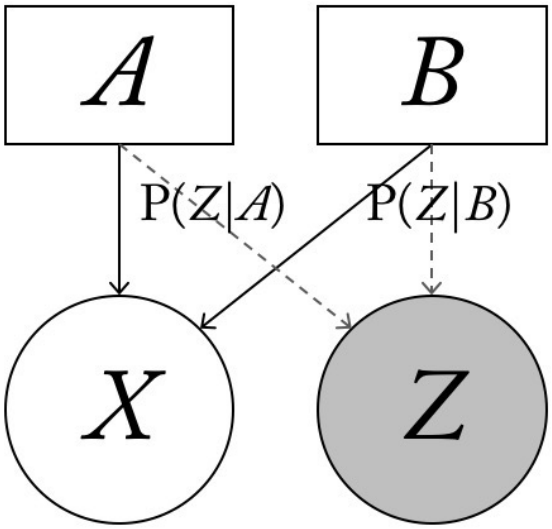

Figure 3. Causal structure where explanations of *X* inform predictions about *Z*.

Chapter 4 argues that, contrary to probability theory, people do not integrate across all of these possibilities when they use one inference as an input to another, but instead use beliefs in a *digital* way that tacitly assigns probabilities of 0 and 1 and allows the computational explosion to be short-circuited. Studies 18 and 19 presented participants with observations *X* (e.g., a pond's rust brown color and excess algae) and two explanations, with explanation *A* better than explanation *B* (e.g., because it is simpler). Participants were asked to predict the probability of event *Z*, where P(*Z*|*A*) and P(*Z*|*B*) differed across conditions (see Figure 3). Participants'

judgments of P(Z) differed based on P($Z|A$) but not P($Z|B$), indicating that they assigned all weight to possibility A when making predictions—that is, they reasoned digitally. This was true even when participants explicitly rated the probabilities of the explanations (Study 20) and when the instructions explicitly assigned B a 30% base rate (Study 21) or posterior probability (Study 22). Only when all conditional probabilities and base rates were given explicitly in the problem did participants at last incorporate uncertainty about the explanations into their predictions (Study 23).

One possible application of digitization is to financial markets, where markets are known to "overreact" to new information (de Bondt & Thaler, 1985): Could overreactions occur because people oscillate digitally between highly optimistic and pessimistic assessments? Studies 24–27 showed that people indeed digitize new information when making financial predictions, ignoring the possibility of lower-probability events in making predictions. Within the sample's range of expertise, investing experience did not attenuate the digitization effect, suggesting that it is a deep-seated cognitive bias.

These biased predictions are problematic, but may nonetheless be the best strategy given the computational explosion from integrating possibilities at each step in a chain of reasoning. Although this strategy will not lead to correct probabilities, by choosing the most likely possibility at each stage of the inference it avoids the computational explosion while assuming the single most plausible possibility. Perhaps for this reason, people use a similar strategy in category-based prediction (Murphy & Ross, 1994) and in vision, which adopts one interpretation of the retinal array at a time (Attneave, 1971). Digitization, like inferred evidence and opponent simplicity heuristics, may be a highly general component of explanatory logic.

Chapter Five: Of the Origins of Sense-Making

Although adults rely on an explanatory logic that supersedes normative Bayesian inference, children may not. If explanatory principles are acquired over a long period of development as inferential "short-cuts," then young children may be less likely to rely on these principles, and thus more likely to reason normatively in explanatory contexts. In contrast, if these principles are present from early in development, then children, like adults, should make some of the same non-normative inferences as adults.

Chapter 5 tests this question using inferred evidence as a case study. Studies 26 and 28 found that 4–8-year-olds, like adults, non-normatively favor narrow latent scope explanations, with little change over development. This bias appears to occur due to the same inferred evidence mechanisms found in adults, since children arbitrarily guess about evidence to inform their explanatory judgments (Study 27). That said, the bias is not blind: Children can override this bias when the prior odds strongly favor a wide latent scope explanation (Study 29).

Together with evidence that 4-year-olds favor simpler explanations (Bonawitz & Lombrozo, 2012), these studies suggest that explanatory logic emerges early. Rather than being abstracted

from repeated instances of normative probabilistic reasoning, it may be that explanatory principles instead provide the scaffolding for probabilistic reasoning.

Chapter Six: Of Social Understanding

Two organizing questions of social psychology are: First, how do human beings place one another into groups and use those groups to judge others? Second, how do we peer into the mind of other people to intuit their thoughts and predict their behavior? The former question is one of *stereotyping*, and the latter of *theory-of-mind*, but both are fundamentally explanatory: What category best explains an individual's traits, and what mental states best explain an individual's behavior? Chapter 6 argues that our explanatory strategies pervade both processes of social understanding.

Previous chapters documented three tools of explanatory logic, and people use all three tools in stereotyping. They use *inferred evidence*: When two social categories are associated with an observed trait (e.g., arrogance), people think an individual is more likely to belong to a category that does not predict unknown traits (Study 30). They use *opponent simplicity heuristics*: While people generally prefer simpler social categorizations (a religious category) over complex, intersectional categorizations (an ethnicity and an occupation category), this preference is weaker when the categories are more loosely associated with the traits (Study 31), just as the simplicity preference in causal explanation is weaker for stochastic systems. And people *digitize*: They act as though an individual with a high probability of belonging to a category *certainly* belongs to that category when predicting their other traits (Study 32). These findings jointly suggest that people view stereotyping as a process of explanation and that this process is governed by the same logic as other explanatory processes.

Many of these same strategies are used in theory-of-mind. When assessing an individual's intention in performing a behavior, people tend to favor intentions that do not make predictions about unknown actions (Study 33). And while people generally favor simple intentions (one goal) over complex intentions (two goals), this preference is weaker when making inferences about individuals with less consistent preferences (Study 34). Thus, theory-of-mind appears to be yet another context in which these explanatory principles are used.

Together, these two lines of work provide converging evidence that we explain the social world using many of the same principles we use in other areas of cognition. A complete picture of social psychology will thus require a detailed understanding of these explanatory practices.

Chapter Seven: Of Choice

Explanations are crucial to choice, because actions often depend on prior inferences. A patient decides on a treatment that matches the disease likeliest to ail her, based on diagnostic tests; an investment banker chooses an asset allocation expected to maximize profits, based on past returns; a consumer chooses the toothpaste likeliest to keep his teeth white, based on

persuasive advertising. To what degree are choices based on explanatory inferences subject to the same errors as the inferences themselves?

Chapter 7 argues that although people rely on many of the same explanatory processes in choice and non-choice contexts, additional cognitive processing in choice contexts can lead to more rational inferences. Study 35 compared causal explanations (e.g., which part of a lawnmower is broken) and choices (e.g., which part of the lawnmower to replace), and found a bias against narrow latent *explanations* but no bias against narrow latent scope *choices*. In a separate condition, participants were asked a causal inference question in a choice *context* where prices were indicated, and here participants were again unbiased, suggesting that a choice *context* rather than choice itself leads to unbiased judgments. Subsequent studies distinguished between possible mechanisms. Whereas choice contexts did not induce an increase in analytical thinking (Study 36), the bias re-emerges in choice contexts when the narrow and wide latent scope explanations are evaluated separately. Thus, choices tend to be unbiased relative to judgments because choice facilitates comparison of alternatives.

These results contribute to debates concerning human rationality. On the one hand, they affirm the generally low opinion of human rationality common in behavioral economics, since choices, depending on context, are either biased (because they match biased explanatory judgments) or incoherent (because they do not match them). However, these results are hopeful in a different sense: The psychological mechanisms underlying choice behavior appear to induce bias-correction processes that can lead people to behave more rationally, even in the absence of economic incentives. Thus, this work not only points to limits on human rationality, but also to the limits on those limits.

Chapter Eight: Of the Likely and the Lovely

The need to make sense of the world drives cognition. Categories allow us to bundle features together to support inference, language allows us to extract meaning from utterances, theory-of-mind allows us to infer others' mental states, causal reasoning allows us to understand present events in terms of the past. Yet, these explanatory problems are not straightforward to solve. While we struggle to make veridical inferences, we must simultaneously face informational limits on evidence, specification limits on priors, and cognitive limits on probabilistic thinking. Given these constraints, it is impressive that human beings are generally talented at explaining relevant aspects of our environment, finding the hidden categories, meanings, mental states, and causes that facilitate prediction and action.

Philosophers have contrasted two views of explanation (Lipton, 2004). On the one hand, Bayesian epistemologists argue that people infer the *likeliest* explanation—the explanation with maximum posterior probability given the evidence. Clearly, this view cannot fully capture the empirical results reported here, where people flagrantly violated the laws of probability in myriad ways across diverse cognitive processes. Other philosophers favor the view that people infer the *loveliest* explanation according to reasoners' intuitions. Exemplifying this view, Einstein said that

he "ask[s] many questions, and when the answer is simple, then God is answering" and Keats claimed that "beauty is truth, truth beauty." While this view does seem to capture the phenomenology of explanation, it does little to help us understand why some explanations seem lovelier than others.

I view these positions as complementary and both necessary to fully understand explanatory cognition: The lovely is a *guide* to the likely. In my view, largely veridical Bayesian computations are *realized* by assessing the explanatory virtues (McGrew, 2003). Because we face limits on information, indeterminacy, and cognition, we need strategies to avoid these problems while making inferences veridical enough to be useful. Proximately, we may view lovely explanations as most satisfying, even as our notions of loveliness are ultimately driven by deeper principles pointing us toward useful explanations, predictions, and choices. People are indeed seduced by lovely explanations, but this may be beneficial much of the time. We often perceive things as beautiful precisely because they are true.

References

Attneave, F. (1971). Multistability in perception. *Scientific American*, *225*, 62–71.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*, 1156–1164.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*, 187–192.

De Bondt, W. F. M., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*, 793–805.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*, 1–35.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

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

Kanizsa, G. (1976). Subjective contours. *Scientific American*, *234*, 48–52.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–535.

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

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *The British Journal for the Philosophy of Science*, *54*, 553–567.

Michotte, A., Thinès, G., & Crabbé, G. (1964). Les complements amodaux des structures perceptives. *Studia Psychologica*. Leuven, Belgium: Publications Universitaires de Louvain.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Peirce, C. S. (1997). *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*. (P. A. Turrisi, Ed.). Albany, NY: State University of New York Press. (Original work published 1903.)

Premack, D., & Premack, A. J. (1994). Levels of causal understanding in chimpanzees and children. *Cognition*, *50*, 347–362.

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

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.