Explanations of Counterfactual Inferences

Brian J. Edwards (Brian.Edwards@u.northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208 USA

Lance J. Rips (Rips@northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208 USA

Abstract

When engaging in counterfactual thought, people must imagine changes to the actual state of the world. In this study, we investigated how people reason about counterfactual scenarios by asking participants to make counterfactual inferences about a series of causal devices (i.e., answer questions such as *If component X had not operated [had failed], would components Y, Z, and W have operated?)* and to explain their reasoning. Participants avoided breaking deterministic causal links (i.e., *W always* causes *X*), but were willing to break probabilistic causal links (i.e., *W sometimes* causes *X*) to keep prior causal events in the same states as in the actual world. Participants' explanations supported this pattern of inferences. When the causal links were deterministic, participants reasoned diagnostically to infer that the states of prior causal events would have been different in the counterfactual world. In contrast, when the links were probabilistic, participants cited the links' unreliability as an explanation for why the states of prior causal events would have been the same as in the actual world. Additionally, participants who were told that a component "had failed" (vs. "had not operated") were more likely to attribute the state of that component to it being "internally broken" and infer that causally upstream components would have operated. Our results suggest that people use their explanation of the antecedent event (the "if" clause) to guide their counterfactual inferences. We discuss the implications of these findings for two rival Bayes-net theories of counterfactual reasoning: Pearl's (2000) and Hiddleston's (2005).

Keywords: Counterfactuals, causation, explanation

Introduction

People often engage in counterfactual reasoning (e.g., *If I hadn't partied the night before the exam, then I would have passed the exam*) to second-guess decisions, attribute credit or blame, and diagnose causal relations (see Byrne, 2005, for a review). Reasoning about counterfactual scenarios such as the preceding example requires imagining changes to the actual state of the world—for instance, imagining a counterfactual world in which I hadn't partied the night before the exam. One of the central issues in the study of counterfactual reasoning is how people re-imagine the world to satisfy the antecedent of a counterfactual scenario. (The *antecedent* is the "if" clause, and we will refer to the "then" clause as the *consequent*.) In particular, what types of events do people keep the same in the actual and counterfactual worlds and what types of events do people change?

One way people might reason about counterfactual scenarios, which we will call *pruning theory*, is by using an intervention to change the state of the antecedent event from the actual state to the counterfactual state and then tracing the consequences of that intervention (Pearl, 2000; see also Woodward, 2003). The intervention severs the causal link between the antecedent and its immediate causes, and as a result of this "graph surgery," the counterfactual states of upstream events would be the same as in the actual world. However, downstream events that are a consequence of the antecedent would change states according to the causal laws governing the system. To illustrate this approach, consider a causal chain $A \rightarrow B \rightarrow C$ and a counterfactual antecedent *If B had not occurred…* (in the actual world, *A*, *B*, and *C* all occurred). A person using pruning theory would intervene on *B* to change the state of *B* from present to absent. Since upstream events (*A*) are unaffected by this intervention, *A* would still have been present in the counterfactual world. But since *C* is an effect of *B*, *B*'s absence would in turn cause *C* to be absent.

Pruning theory might appeal to reasoners in two ways. First, by keeping all the events that are causally prior to the antecedent in the same states as in the actual world, pruning theory creates a counterfactual world that is maximally similar to the actual world with respect to these prior events. Second, the pruning approach makes counterfactual thinking computationally easy. The strategy of always keeping prior events in their original states allows reasoners to avoid the cognitively challenging process of reasoning backwards to determine the counterfactual states of upstream causes.

However, other researchers have questioned whether the type of change pruning theory proposes is necessarily the most reasonable way to modify the causal system in the counterfactual situation (e.g., Hiddleston, 2005). One criticism of pruning theory is that it is very disruptive to the structure of a causal system and can require reasoners to violate causal laws. Consider a deterministic causal system in which *A*, without exception, always causes *B*. In this setting, one might be reluctant to imagine a counterfactual world in which *A* occurred, but *B* did not occur (e.g., in answering the question *If B had not occurred, would A have occurred?*). Thus, when reasoning about this counterfactual scenario, one might be more likely to infer that the reason *B* did not occur was that *A* did not occur, and the absence of *A* caused *B* to be absent too (Hiddleston, 2005). We will call this alternative *minimal-network theory*. When the causal links are probabilistic (i.e., *A sometimes* causes *B*), however, minimal-network theory proposes that *A* might or might not have occurred, since either possibility is "legal" in accordance with the system's causal laws.

Table 1 compares the predictions of pruning theory and minimal-network theory for a device in which component A's operating usually causes component B to operate and component B's operating always causes component C to operate (at present, all three components are operating). The device's structure is illustrated as follows:

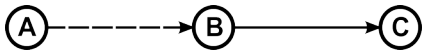


Table 1: Comparison of Pruning Theory and Minimal-network Theory

	If component B had not operated, would component A have operated?	If component C had not operated, would component B have operated?
Pruning Theory	Yes	Yes
Minimal-network Theory	Maybe	No

Previous empirical work has explored whether people's counterfactual inferences are consistent with either of these two theories of counterfactual reasoning. In one experiment, Sloman and Lagnado (2005) presented people with causal information about a simple rocket-ship device with the causal structure $A \rightarrow B$ and asked them a variety of counterfactual questions. Sloman and Lagnado found evidence that people engaged in pruning when they were told that a component was *prevented* from operating, but not when told that the component was *observed* not to have operated. However, subtle differences in wording across their experiments led to significantly different patterns of counterfactual inferences, making it difficult to generalize from the data. In another study, Rips (2010) asked people counterfactual questions about three- and four-component mechanical devices. Although participants' counterfactual inferences did not provide strong support for either pruning theory or minimal-network theory, their inferences were more closely aligned with minimal-network theory (see also Dehghani, Iliev, & Kaufmann, 2012).

In the two experiments in this study, we presented participants with counterfactual questions for which pruning theory and minimal-network theory make different predictions. The wording of these questions was manipulated across two between-subjects conditions. One group of participants was told that a component of a mechanical device "had not operated," and another group was told that the component "had failed" (e.g., *If Component B had not operated/had failed…*). The neutral "had not operated" wording does not suggest a particular explanation for the state of the component; however, the "had failed" wording suggests an explanation that is local to the component (e.g., the component is internally broken). Thus, we predict that participants in the *not operated* and *failed* conditions will make different counterfactual inferences about the operating states of the other components. Specifically, we predict that participants in the *not operated* condition will reason diagnostically about the states of the other components based on the device's causal structure, consistent with minimal-network theory. In contrast, we predict that participants in the *failed* condition will reason that since the antecedent component is broken, its operating state is not diagnostic of the states of the other components. Thus, participants will break the causal links between the antecedent and its causes and infer that causally prior components would have operated in the counterfactual situation, consistent with pruning theory. In addition to examining participants' inferences about which components would and would not have operated in the counterfactual situation, we analyzed participants' explanations of their reasoning. In Experiment 1, we analyzed people's explanations of why they thought the non-antecedent components would or would not have operated. In Experiment 2, we analyzed people's explanations of why the antecedent event would have occurred.

Experiment 1

Participants in this experiment received a series of problems about a set of eight hypothetical devices, each with four components. For each device, they answered counterfactual questions of the form *If component X had not operated [had failed], would components Y, Z, and W have operated?* and provided explanations justifying their reasoning.

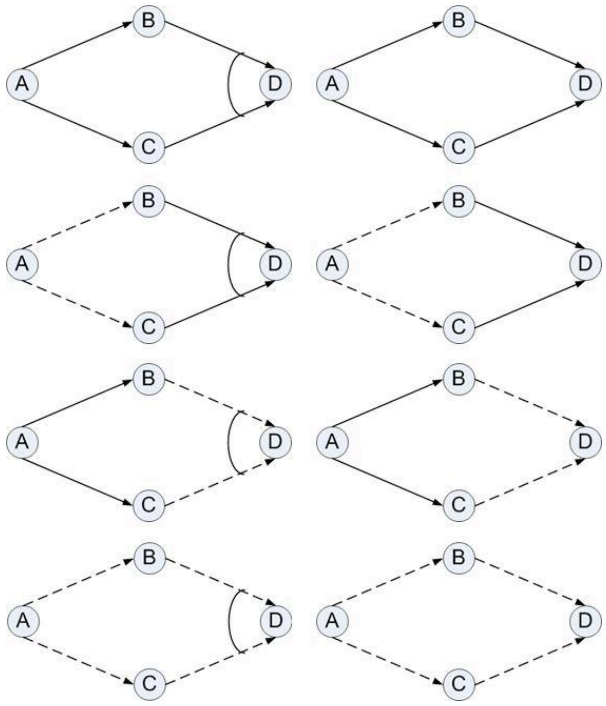
Method

Materials. The questionnaire booklets contained three pages of instructions followed by 24 pages of questions. The instructions explained the experimental task and told participants how to interpret the diagrams of the causal devices on the following pages. Each question page contained a written description of how a device operated (e.g., Component A's operating always causes component B to operate, etc.), which was accompanied by the corresponding diagram in Figure 1.

As shown in Figure 1, there were eight different causal devices, all of which had "diamond" structures. The devices varied in whether the causal links between components were deterministic (solid lines in Figure 1) or probabilistic (dashed lines), and whether components B and C had to operate together to cause D to operate (arc connecting links in Figure 1) or could independently cause component D to operate (no arc). The order of the devices was counterbalanced across participants. We used devices with diamond structures for two reasons. First, previous causal reasoning studies have used diamond structures and have found that people make accurate causal inferences about these systems (Meder, Hagmayer, & Waldmann, 2008, 2009). Second, pruning theory and minimal-network theory make different predictions for many of the counterfactual questions about these devices.

After learning how each device works, participants were told the device's current operating state, which was always that "at present, components A, B, C, and D are all operating." Next, participants were asked a counterfactual question about the device, such as *If component B had not operated, would components A, C, and D have operated?* For each of the eight devices, participants answered three counterfactual questions, one question each with A, B, and D as the antecedent component. Since the devices were symmetric with respect to components B and C, we did not ask a separate question in which C was the antecedent. The order of the antecedent components for these questions (ABD vs. DBA) was balanced across participants.

Figure 1: Causal Devices Used in Experiments 1 and 2



In this figure, solid arrows indicate deterministic links (e.g., A always causes B) and dashed arrows indicate probabilistic links (e.g., A usually causes B). All causal relationships are in the direction shown by the arrows. The arcs indicate that component B and component C operating together cause component D to operate, but component B or component C operating alone never causes component D to operate (jointly caused devices). The absence of an arc indicates that component B or component C operating alone causes component D to operate (separately caused devices).

Participants were randomly assigned to one of two experimental conditions. In the *not operated* condition, participants learned that the antecedent component (component B in the preceding example) "had not operated." In the *failed* condition, participants learned that the antecedent component "had failed."

For each counterfactual question, participants indicated which of the three non-antecedent components would have operated in the counterfactual state. For each component, participants could say that the component (1) would have operated, (2) would not have operated, or (3) might or might not have operated. To gain insight into how participants were reasoning about the counterfactual questions, participants also indicated the order in which they reasoned about the non-antecedent components. After making these inferences, participants justified their answers by responding to the prompt "Please explain why you answered in the way you did."

Procedure. Participants received the questionnaire booklet from the experimenter and answered the questions at their own pace. The experiment took approximately 30 minutes to complete.

Participants. Participants were 32 undergraduate students at Northwestern University. Participants received course credit for their participation.

Results and Discussion

We analyzed participants' answers to the counterfactual questions (e.g., *If component B had not operated, would component A have operated?*) to see if their inferences were consistent with minimal-network theory or pruning theory. Responses of "would have operated" were scored as +1, responses of "would not have operated" were scored as -1, and responses of "might or might not have operated" were scored as 0. The mean score for participants was higher in the *failed* condition (M = -0.14) than in the *not operated* condition (M = -0.43), $F(1, 32) = 7.07$, $MSe = 7.29$, $p = .01$.

In two cases, pruning theory and minimal-network theory make the same predictions: (1) when component A was the antecedent, and (2) for the devices in which components B and C must both operate in order for component D to operate (jointly caused devices), when component B was the antecedent and component D was the consequent. In case (1), both theories say that components B, C, and D would all not have operated, and in case (2), both theories say that component D would not have operated. For all the other counterfactual questions, pruning theory predicts that the consequent component definitely would have operated (producing positive scores), whereas minimal-network theory predicts that the consequent component either (a) definitely would not have operated or (b) might or might not have operated (producing negative or 0 scores respectively).

When we restricted our analysis to the cases in which pruning theory and minimal-network theory make different predictions, the mean score for participants in the *failed* condition was 0.17 and the mean score for participants in the *not operated* condition was -0.27. As was the case with the entire data set, the difference between conditions was significant, $F(1, 32) = 11.96$, $MSe = 6.27$, $p = .002$. The mean score for the *not operated* condition was significantly less than 0, $t(17) = -4.50$, $p < .001$; however, the mean score

for the *failed* condition was not significantly different from 0, $t(17) = 1.68$, n.s.

Next, we examined the serial order (1, 2, or 3) in which participants reasoned about the three non-antecedent components. The most interesting case is the one in which component B was the antecedent since participants could work their way downstream (i.e., reason about component D first) or upstream (i.e., reason about component A first). Most participants (69%) started upstream, reasoning about component A before component D (Binomial test, $p < .001$). The mean serial position for component A was 1.44, whereas the mean position for component D was 2.32. The order in which participants reasoned about the components did not differ across the *failed* and *not operated* conditions,.

We also examined participants' explanations of their counterfactual reasoning to see if the explanations were consistent with pruning theory or minimal-network theory. We classified explanations in two ways.

(1) Explanations were coded as *causal backtracking* if participants used the state of the antecedent component to reason diagnostically about the states of upstream components. A sample causal-backtracking explanation was "If B wasn't operating that would mean A wasn't working since A always causes B." Causal-backtracking explanations are consistent with minimal-network theory.

(2) Explanations were coded as *causes are independent of effects* if they suggested that the states of upstream "cause" components are not affected by the states of downstream "effect" components. A sample explanation was "Neither A, B, nor C are dependent on D so they all will have operated." Such an explanation is consistent with pruning theory.

Notice that these three types of explanations are only applicable when there are components that are causally upstream of the antecedent component. Thus, we restricted the following analyses to the counterfactual questions in which B or D was the antecedent. The data were coded by a person who was unfamiliar with the experimental hypotheses, and 25% of the data were coded independently by a second coder. Inter-coder reliability was 90%.

Participants in the *not operated* condition were significantly more likely to provide "causal-backtracking" explanations than participants in the *failed* condition (65% vs. 32% respectively, $F(1,24) = 12.9$, $MSe = 16.4$, $p = .001$). In contrast, participants in the *failed* condition were significantly more likely to provide "causes are independent of effects" explanations than participants in the *not operated* condition (25% vs. 9% respectively, $F(1, 21) = 5.57$, $MSe = 3.90$, $p = .03$). Participants in the *not operated* condition were significantly more likely to provide "causal-backtracking" explanations than "causes are independent of effects" explanations ($t(14) = 6.33$, $p < .001$); however, participants in the *failed* condition did not significantly prefer either type of explanation.

In sum, participants in the *not operated* and *failed* conditions differed in their counterfactual inferences. Participants in the *not operated* condition had a stronger tendency to say that non-antecedent components would not have operated than participants in the *failed* condition, and they made inferences that were better predicted by minimal-network theory. The analysis of participants' explanations also showed that most participants in the *not operated* condition used causal backtracking to diagnose the counterfactual operating states of upstream components. In contrast, participants in the *failed* condition were more likely than participants in the *not operated* condition to say that the operating states of upstream components were independent of, and could not be diagnosed from the state of the antecedent.

Experiment 2

The pattern of inferences and reasoning strategies in Experiment 1 suggests that participants in the *not operated* and *failed* conditions may have generated different explanations for why the antecedent component had not operated. We therefore performed a second experiment to investigate the possible relationship between participants' explanations of why the antecedent component had not operated and their counterfactual inferences.

Method

The experiment contained two parts, an inference task and an explanation task. The same eight causal devices from Experiment 1 were used in Experiment 2 (see Figure 1). As in Experiment 1, participants were randomly assigned to either the *not operated* condition or the *failed* condition.

Materials.

Inference task: The inference task was identical to Experiment 1 except that participants did not provide explanations of their counterfactual inferences during this part of the experiment.

Explanation task: In the explanation task, participants described why the *antecedent* component had not operated. Note that this is a different type of explanation than the ones participants provided in Experiment 1; in Experiment 1, participants explained why the *non-antecedent* components would or would not have operated. The explanation-task booklet included three pages of instructions followed by 24 pages of questions. As in the inference task and Experiment 1, participants received information about how the causal devices work and told that "at present, components A, B, C, and D are all operating." Participants in the *not operated* condition were asked questions of the form *If component X had not operated, which of the following would best explain why?* Participants in the *failed* condition were asked a question that was identical except that "not operated" was replaced by "failed." For each device, participants answered this question for each of components A, B, and D as the antecedent. For each participant, the order of the devices, and within each device, the order of the antecedent components, was the same in the inference and explanation tasks.

When component B was the antecedent, participants selected an explanation from the following list:

(1) Component B was internally broken.
(2) Factors external to the device prevented component B from operating.
(3) Component B operates unreliably, and component B just didn't operate this time.
(4) Component A did not operate, which in turn caused component B not to operate.
(5) Component A operated, but component B just didn't operate this time because the connection between component A and component B is unreliable.
(6) Component A operated, but the connection between component A and component B was broken.

The list of explanations was similar when component D was the antecedent, except that "component D" was substituted for "component B" and "component B and/or[1] component C" was substituted for "component A." When component A was the antecedent, only the first three answer choices were included since component A's operation is not caused by other components. The order of the answer choices (above order vs. reverse order) was balanced across participants.

After choosing an explanation, participants rated their confidence on a 0-9 scale with one-point increments, where 0 = "not at all confident" and 9 = "extremely confident."

Procedure. Half of the participants completed the inference task followed by the explanation task and the remaining participants completed the explanation task followed by the inference task. Each task took approximately 20 minutes with the entire experiment taking approximately 40 minutes.

Participants. Participants were 32 undergraduate students at Northwestern University who had not participated in Experiment 1. Participants received course credit for their cooperation.

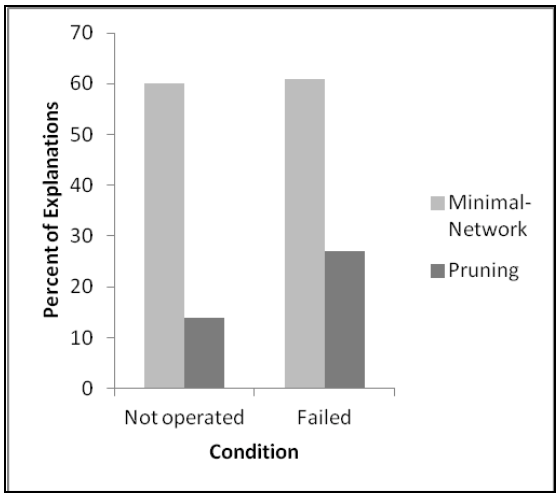
Results and Discussion

Inference Task. The inference task replicated the findings of Experiment 1. The mean score for participants in the *failed* condition was significantly higher than for participants in the *not operated* condition. This was true for all counterfactual questions (M = -0.16 vs. M = -0.48 respectively, $F(1,30) = 14.47$, $MSe = 4.14$, $p < .001$) and for the subset of counterfactual questions for which pruning theory and minimal-network theory make different predictions ($M = 0.27$ vs. $M = -0.25$ respectively, $F(1, 30) = 19.27$, $MSe = 4.94$, $p < .001$).

Explanation Task. Participants' explanations were coded as consistent with pruning theory, consistent with minimal-network theory, or consistent with neither theory. Explanations 1, 2, and 6 (see Method section) were classified as pruning explanations. When the links between the antecedent and its causes were deterministic, explanation 4 was classified as a minimal-network explanation. When the links between the antecedent and its causes were probabilistic, explanation 5 was classified as a minimal-network explanation. All other responses were classified as "other." Since neither explanation 4 nor explanation 5 (the two possible minimal-network explanations) is applicable when component A was the antecedent, the following analyses were conducted only for the counterfactual questions in which component B or D was the antecedent.

Participants were significantly more likely to choose minimal-network explanations than pruning explanations (61% vs. 21% respectively, $t(31) = 5.31$, $p < .001$). This pattern was observed in both the *not operated* (60% vs. 14% respectively, $t(15) = 4.92$, $p < .001$) and *failed* conditions (61% vs. 27% respectively, $t(15) = 2.85$, $p = .01$). Notice that participants in the *failed* condition were significantly more likely to choose pruning explanations than participants in the *not operated* condition ($F(1,29) = 4.56$, $MSe = 3.04$, $p = .04$). The results are shown in Figure 2.

Figure 2: Percent of Minimal-Network and Pruning Explanations by Condition



Interestingly, when the causal links between the antecedent and its causes were probabilistic, participants in both conditions were significantly more likely to choose a minimal-network explanation (e.g., Component A operated, but component B just didn't operate this time because the connection between component A and component B is unreliable) than a pruning explanation (e.g., Component B was internally broken; Factors external to the device prevented component B from operating; Component A operated, but the connection between component A and component B was broken), (*Not operated* condition: $t(15) = 5.81$, $p < .001$, *Failed* condition: $t(15) = 5.00$, $p < .001$). All these explanations (both the pruning and minimal-network explanations) imply, and in some cases state explicitly, that causally upstream components would have operated. Even

[1] If either component B or component C operating alone could cause component D to operate, the "and" wording was used. Otherwise, the "or" wording was used.

though this counterfactual state is consistent with both pruning theory and minimal-network theory, participants in both conditions preferred minimal-network explanations.

As in Experiment 1, minimal-network theory better explained the inferences of participants in the *not operated* condition compared to pruning theory. While participants in the *failed* condition were more likely than participants in the *not operated* condition to say that non-antecedent components would have operated, participants in both conditions preferred minimal-network explanations over pruning explanations. Thus, Experiment 2 suggests that minimal-network theory might provide a starting point for a good psychological theory of counterfactual reasoning.

General Discussion

In the two experiments in this paper, we examined (1) participants' counterfactual inferences about the states of variables in a causal system and (2) participants' explanations of their reasoning. Alternative theories of counterfactual reasoning such as pruning theory and minimal-network theory make different predictions about how people should modify (or preserve) the system's causal structure when reasoning about a counterfactual scenario.

A defining characteristic of pruning theory is the proposal that people treat counterfactuals as interventions. Under this account, people should simulate the counterfactual state by intervening on the causal system, and we would expect them to break both probabilistic and deterministic causal links and say that upstream components would have operated. Furthermore, they should endorse an interventionist explanation for the counterfactual state of the antecedent component, such as "factors external to the device prevented the antecedent component from operating."

Our data provide evidence against this hypothesis. Participants in the neutrally worded *not operated* condition made counterfactual inferences that preserved deterministic causal relationships between components' operating states. When the causal links between the antecedent component and its causes were deterministic, participants inferred that the antecedent component's causes would not have operated, which in turn caused the antecedent component not to operate. However, when the causal links were probabilistic, participants inferred that the antecedent component's causes would have operated, but the antecedent component would not have operated because the links were unreliable. These inferences and explanations are consistent with minimal-network theory, which proposes that people should prefer "legal" counterfactual states that preserve the system's (deterministic) causal laws, but they are inconsistent with pruning theory.

We also found that participants in the *not operated* and *failed* conditions reasoned differently about the counterfactual scenarios. The *failed* wording suggested to participants that the antecedent component was internally broken. Accordingly, these participants modified the devices' causal structure by breaking the causal links between the antecedent and its causes, and they inferred that upstream components would have operated. Other studies that have varied the wording of counterfactual questions have found similar effects (Sloman & Lagnado, 2005).

Each type of wording supports a particular (and different) explanation for the counterfactual antecedent. The differences in participants' explanations across conditions suggest that these explanations may in turn shape participants' counterfactual inferences. Hempel (1965) famously proposed that causal explanations support predictive inferences, and our data suggest such a connection between explanation and inference in counterfactual reasoning (Goodman, 1955). Specifically, we propose that when engaging in counterfactual reasoning, people integrate their explanation of the counterfactual antecedent with their knowledge of the system's causal structure to infer the system's counterfactual state.

Acknowledgements

We thank Ben Rottman, Steven Sloman, and members of the Northwestern University Higher Level Cognition Laboratory for their valuable feedback on these experiments. We thank Samantha Thompson and Joanna Westerfield for their research assistance. The research was supported by an NSF graduate research fellowship (BJE) and IES Grant R305A080341 (LJR).

References

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind and Language.*

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

Hempel, C. G. (1965). Aspects of scientific explanation. In C. G. Hempel, *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 331-496). New York: Free Press.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous, 39*, 632-657.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15*, 75-80.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition, 37*, 249-264.

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

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

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

Woodward, J. (2003). *Making things happen*: *A theory of causal explanation*. New York: Oxford University Press.