A unified theory of counterfactual reasoning

Christopher G. Lucas
cglucas@cmu.edu
Department of Psychology
Carnegie Mellon University

Charles Kemp
ckemp@cmu.edu
Department of Psychology
Carnegie Mellon University

Abstract

A successful theory of causal reasoning should be able to account for inferences about counterfactual scenarios. Pearl (2000) has developed a formal account of causal reasoning that has been highly influential but that suffers from at least two limitations as an account of counterfactual reasoning: it does not distinguish between counterfactual observations and counterfactual interventions, and it does not accommodate backtracking counterfactuals. We present an extension of Pearl's account that overcomes both limitations. Our model provides a unified treatment of counterfactual interventions and backtracking counterfactuals, and we show that it accounts for data collected by Sloman and Lagnado (2005) and Rips (2010).

In addition to reasoning about actual states of affairs, humans find it natural to reason about what might have been. A doctor may ask "if Alice had not been treated with the experimental drug, would she have survived?" and a parent might tell a child that "if you had been paying attention, you wouldn't have gotten hurt." Researchers from several disciplines have developed formal models of counterfactual reasoning, and recent empirical studies have evaluated the psychological merits of some of these models (Rips, 2010; Dehghani, Iliev, & Kaufmann, 2012). This paper describes a new model of counterfactual reasoning and evaluates it using data sets from the psychological literature.

The problems that we consider can be illustrated using a causal chain over three variables (Figure 1a). For example, suppose that A, B, and C are variables that indicate whether three transponders are active. Transponder A is active about half of the time, and whenever it is active it tends to activate B, which in turn tends to activate C. Suppose that we observe on a certain occasion that all three transponders are active. We can now ask counterfactual questions such as "if B had not been active, would C have been active?"

The formal approach that we present is inspired by the work of Pearl (2000), who developed a model of counterfactual reasoning that we refer to as the *modifiable structural model*, or MSM for short. The MSM assumes that the causal system in question is a *functional causal model*, where *exogenous* variables are introduced if necessary so that the variables of primary interest are deterministic functions of their parents. For example, the system in Figure 1a may be represented more precisely by adding exogenous variables U_A, U_B and U_C such that U_A determines whether or not node A is active, and U_B and U_C capture factors such as atmospheric conditions that determine whether the links in the chain operate successfully. Suppose now that A, B and C are all observed to be active, and that we want to know whether C would be active if B were not active. The MSM addresses this question by using the observations in the actual world to update

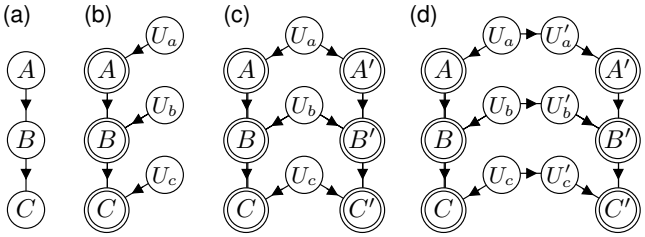

Figure 1: (a) A causal chain in which A causes B, which causes C. (b) A functional causal model that captures the causal chain in (a). Exogenous variables U_a, U_b and U_c have been added, and nodes with double edges are deterministic functions of their parents. (c) A twin network where A', B' and C' represent counterfactual values of A, B, C. (d) An augmented twin network which allows for the possibility that the exogenous variables take different values in the counterfactual scenario.

prior beliefs about the status of the exogenous variables, then *modifying* the resulting causal model to reflect a counterfactual intervention where B is forced to be inactive. Inferences about any other variables can be computed using the modified causal model—for example, if B were inactive, then C would probably also be inactive. Several psychological studies of counterfactual reasoning have evaluated the predictions of the MSM and have found that some counterfactual inferences do appear to be treated as inferences about counterfactual interventions (Sloman & Lagnado, 2005; Kemp, Shafto, & Tenenbaum, 2012).

The approach we present builds on the key ideas behind the MSM, and we refer to it as the *doubly-modifiable structural model* or DMSM for short. Like the MSM, the DMSM works with causal systems that are represented using functional causal models and allows these systems to be modified via counterfactual interventions. In addition, however, the DMSM permits a second kind of modification where exogenous variables are altered not because of a counterfactual intervention, but simply because the counterfactual world might have turned out differently from the real world. An important consequence of this difference is that the DMSM alone accounts for backtracking counterfactuals. For example, suppose again that all three transponders in A were observed to be active, and we are asked to decide whether A would be active if B were inactive. The DMSM allows for the possibility that variables upstream of B might explain the counterfactual premise that B is inactive, and thus predicts that A is likely to be inactive. The MSM, however, can only reason about

the downstream consequences of a counterfactual intervention that renders B inactive.

The inability of the MSM to deal with backtracking counterfactuals is widely acknowledged, and Hiddleston (2005) has developed a *minimal networks* model that overcomes this limitation. Rips (2010) has recently evaluated the psychological merits of Hiddleston's account, and reports that Hiddleston's model can account for human inferences about backtracking counterfactuals when supplemented with additional psychological assumptions. Although this variant of Hiddleston's approach accounts relatively well for the data collected by Rips, the minimal networks model is not well suited for reasoning about counterfactual interventions, and does not account for empirical data suggesting that counterfactual inferences are sometimes treated as inferences about counterfactual interventions.

At present, then, the psychological literature on counterfactual reasoning is fragmented. The MSM provides an elegant account of reasoning about counterfactual interventions but does not account for inferences about backtracking counterfactuals. The minimal networks model can handle backtracking counterfactuals, but does not give a clear account of inferences about counterfactual interventions. In contrast, the DMSM accommodates both counterfactual interventions and backtracking counterfactuals, and we will show that it accounts for previously-published experiments that explore both kinds of inferences. As we discuss towards the end of the paper, the DMSM is not a complete account of counterfactual reasoning, but we believe that it comes closer to this goal than any previous model.

The Modifiable Structural Model (MSM)

The MSM was introduced informally above, and we now describe how the predictions of this model can be computed by constructing and manipulating a twin network (Pearl, 2000). The first step is to specify a functional causal model such as the example in Figure 1b that captures the causal system under consideration. Functional causal models are described in detail by Pearl (2000), but for our purposes, their most important feature is that they represent noise or randomness using unobserved exogenous variables, rather than inherently stochastic relationships. This functional model is converted into the twin network in Figure 1c by adding nodes A', B' and C' that represent counterfactual versions of A, B, and C. The counterfactual nodes A', B' and C' and the original nodes A, B, and C have the same exogenous variables as parents, which captures the idea that the causal mechanisms in the counterfactual world are identical to the causal mechanisms in the actual world. Given the twin network, a counterfactual premise can be captured using an intervention that fixes the value of one of the counterfactual variables. For example, suppose again that A, B and C are all active and we are asked about a scenario where B is inactive. The counterfactual premise is captured using graph manipulation to modify the twin network. In other words, we set B' to 0, and remove all arrows between B' and its parents to reflect the fact that B' was fixed by an intervention instead of being brought about by U_b and A'. We can now use the manipulated twin network to compute predictions about the other counterfactual variables. Because A must have been caused by U_A and U_A also causes A', the MSM infers that A' is active. Because B' is inactive, the MSM infers that C' is also inactive.

Two aspects of the MSM are worth emphasizing for comparison with the DMSM described in the next section. First, the MSM handles all counterfactual queries by reasoning about counterfactual interventions. The model therefore does not distinguish between counterfactual interventions ("imagine that someone had disabled transponder B") and counterfactual observations ("imagine that you had observed that B was inactive"). Second, the MSM cannot make inferences about backtracking counterfactuals. If asked to imagine that B were inactive, the MSM fixes the status of B' by means of an intervention and therefore cannot reason about upstream variables such as A' which may explain the inactivity of B'.

The Doubly-Modifiable Structural Model (DMSM)

Just as the MSM can be characterized in terms of computations over a twin network, the DMSM can be characterized in terms of computations over an *augmented twin network*. The augmented twin network for the three element chain is shown in Figure 1d. Note that the network includes nodes for counterfactual versions of the exogenous variables U_A, U_B and U_C in addition to nodes for counterfactual versions of A, B and C. The value of each counterfactual exogenous variable is either copied across from the corresponding real-world variable or generated from the same distribution as the corresponding real-world variable. More precisely, if $P_i(\cdot)$ is the prior distribution on exogenous variable U_i, the value of U_i' is drawn from the distribution

$$P(U_i'|U_i) = s\delta(U_i) + (1-s)P_i(U_i)$$

where $\delta(U_i)$ is a delta distribution that takes value 0 at every point except $U_i' = U_i$ and s is a stability parameter where $0 \leq s \leq 1$. If $s = 1$, then the exogenous variables are perfectly stable, which means that $U_i' = U_i$ for all i and that the DMSM is equivalent to the MSM[1]. If $s = 0$, then the exogenous variables are maximally unstable, and the values of U_i' and U_i are independently drawn from the distribution $P_i(\cdot)$. We will refer to this special case as the USM, or "unattached structural model" because setting $s = 0$ decouples the counterfactual nodes from the actual nodes, meaning that the model effectively discards all observations of the actual world.

We propose that people are sensitive to both the true state of the world and base rate information, and therefore hypothesize that the judgments of most individuals reflect stability values between 0 and 1. A second hypothesis is that some individuals always use $s = 0$ and others always use $s = 1$. A third hypothesis is that each individual uses $s = 0$ in some

[1] A stability of 1 for counterfactual observations can lead to mutually incompatible or impossible states, so we assume that all counterfactual premises are treated as interventions when s = 1.

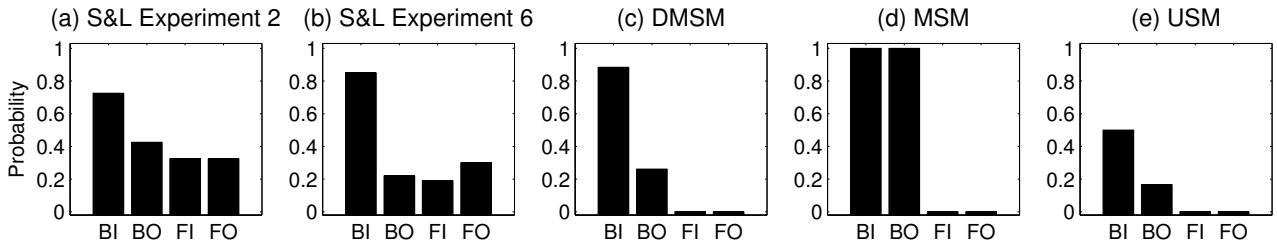

Figure 2: Human judgments and predictions of the DMSM, MSM and USM models for Sloman and Lagnado's Experiments 2 and 6. The four bars in each plot show inferences about backtracking (B) or forward (F) counterfactuals that were either interventional (I) or observational (O).

contexts and $s = 1$ in other contexts. We return to these hypotheses later and describe some preliminary evidence that supports the first hypothesis. Different individuals may use different settings of s, but for all analyses we set $s = 0.77$ which is the value that maximizes model performance across the entire set of studies reported by Rips (2010).

The augmented twin network can be used to address two kinds of counterfactual queries. Queries about counterfactual *interventions* are addressed by manipulating the network in the standard way—for example, counterfactual interventions on B are captured by fixing the value of B' and removing all arrows between B' and its parents. Queries about counterfactual *observations* are carried out by reasoning over the unmanipulated network. For example, if A, B and C are observed to be active and we want to reason about a case where B is observed to be inactive, we set $A = B = C = 1$ and $B' = 0$ and can subsequently compute the posterior distribution induced on any other node. For instance, if the stability parameter is less than one then the posterior distribution $P(A'|A = B = C = 1, B' = 0)$ will indicate that A' is relatively likely to take value 0.

Counterfactual interventions vs. observations

A key conceptual difference between the MSM and the DMSM is that the DMSM allows for counterfactual observations and counterfactual interventions, but the MSM treats all counterfactual queries in terms of interventions. To the best of our knowledge, two experiments reported by Sloman and Lagnado (2005) are the only psychological studies that contrast counterfactual observations and counterfactual interventions. This section compares the predictions of the DMSM with the results of these experiments.

Experiment 2 of Sloman and Lagnado (2005) considers a three node chain where A causes B and B causes C. The experiment included three different cover stories—one scenario involved a rocket ship, and a second involved a causal chain where smoking causes cancer which causes hospitalization. The third involved abstract events A, B, and C and participants were told that "when A happens, it causes B most of the time" and "when B happens, it causes C most of the time." In all cases, participants were told that A and C happened and were asked to make counterfactual inferences about a situation where B did not happen. The counterfactual questions asked about both counterfactual interventions and counterfactual observations. In the abstract scenario, the backwards intervention (BI) question stated that "someone intervened directly on B, preventing it from happening," and asked participants to rate the probability that A would have happened. The forward intervention (FI) question was similar except that it asked participants to rate the probability that C would have happened. The backwards and forwards observation questions (BO and FO) asked participants to rate the probability that A and C "would have happened if we observed that B did not happen."

Average human responses are shown in Figure 2a. Responses were originally provided on a 1 to 5 scale, but we map them to probabilities for comparison with model predictions. Following Sloman and Lagnado (2005), responses are collapsed across the three different cover stories. Both the intervention and observation questions produce the forward inference that C is unlikely to occur. The two kinds of questions, however, lead to different backward inferences about A. Participants tend to infer that A would still have occurred if a counterfactual intervention had prevented B, but find it less likely that A would have occurred if B had been *observed* not to occur.

Experiment 6 of Sloman and Lagnado (2005) is similar in structure but involves a two node chain rather than a three node chain. The cover story described a rocket ship with two components where "movement of component A causes component B to move." Participants were informed that both components are moving and asked to reason about counterfactual cases where either A or B was not moving. The counterfactual intervention questions were of the form "suppose component A were prevented from moving, would component B still be moving?" The counterfactual observation questions were of the form "suppose component A were observed to be not moving, would component B still be moving?"

The proportion of participants who responded "yes" to each question is shown in Figure 2b. As for Experiment 2, forward inferences about B given A are similar regardless of whether A is prevented from moving or simply observed not to move. Backward inferences about A given B again reveal a difference between counterfactual observations and counterfactual interventions. As for Experiment 2, participants tend

to infer that A would still be moving if B were prevented from moving, but are less likely to infer that A would be moving if B were observed not to move.

We generated model predictions for the two experiments by making the simplest possible assumptions about the parameters in each causal structure. The base rate for node A was set to 0.5, and the strength of each causal link was set to 0.8 to capture the fact that causes produce their effects "most of the time." To keep the analysis simple we assumed that nodes B and C had no background causes. Given these assumptions, predictions of the DMSM and the MSM are shown in Figures 2c and 2d. The DMSM accounts for the result that counterfactual interventions and counterfactual observations are treated differently. The MSM accounts for human responses to the intervention questions, but makes identical predictions about responses to the observation questions.

Although the DMSM performs better than the MSM, the quantitative predictions of the DMSM depart from human inferences in some cases. For example, humans give non-zero responses to the forward questions in Experiment 6, but the DMSM infers that component B is definitely not moving if component A is not moving. Including a background cause of B would allow the DMSM to match the human responses to the forward questions more closely.

The MSM can be viewed as a special case of the DMSM where the stability parameter s is equal to 1, and the USM model in Figure 2e is the special case where $s = 0$. The USM distinguishes between counterfactual observations and interventions, but its inferences in the BI case are not shaped by the observation that event A occurred in the real world. As a result, the model falls back on the baseline probability that A occurs, and does not account for the human inference that A probably occurred in the counterfactual scenario. Note, however, that these data do not permit us to distinguish between the DMSM and a mixture of MSM and USM strategies. The analysis in the next section will partially address this issue.

Backtracking counterfactuals

The previous section suggested that the DMSM improves on the MSM by distinguishing between counterfactual interventions and counterfactual observations. One important consequence of this distinction is that the DMSM alone is able to handle backtracking counterfactuals, or queries where a reasoner must think about causes that might be responsible for a counterfactual premise. Rips (2010) has carried out an extensive psychological study of backtracking counterfactuals, and this section argues that the DMSM accounts for Rips' data about as well as the minimal network model that he advocates. Dehghani et al. (2012) have also developed a theory that handles backtracking counterfactuals, and have presented some data in support of their theory. They do not describe a fully-specified computational model, but we were able to implement a model that we believe is consistent with their core assumptions. This model, however, did not account for Rips' data as well as the minimal network model, and we therefore focus here on comparing the DMSM with the minimal network model.

Experiment 3 in Rips (2010) asked participants to reason about four causal systems shown at the top of Figure 3. Each system includes components L and H which cause component C to operate. The systems in Figure 3a include two cases where the operation of C is *jointly caused*: arcs between edges in Figure 3a indicate that L and H operate together to cause C to operate. The remaining systems are cases where the operation of C can be *separately caused* by either L or H. Two of the systems include probabilistic causal relationships shown as dashed arrows. For example, the probabilistic jointly caused system was described as a system where component L's operating and component H's operating together usually cause component C to operate. The remaining two systems include deterministic causal relationships. For each of the four systems, base rates for causes L and H were provided. Participants were told that L operates 5% of the time, and that H operates 95% of the time.

In each case participants were told that components L, H, and C "are all operating." They then responded to the question "if component C were not operating, would component L be operating?" and answered a similar question with respect to component H. The red points in Figure 3b show the proportion of participants who said yes to each question. For the deterministic separately caused system, most participants inferred that neither L nor H would be operating in the counterfactual scenario. For all remaining systems, participants tended to infer that the cause with higher base rate would be operating.

Predictions for four models are shown in Figure 3 using black lines. To generate these predictions, we again assumed that the probabilistic causal relationships had a strength of 0.8. The experimental materials did not explicitly specify whether each counterfactual scenario involved an intervention or an observation, and the predictions for the DMSM and the USM are based on counterfactual observations.

The DMSM predicts that average responses for the deterministic separately caused system should be low, and accounts for the base rate effects observed for all other systems. In contrast, the MSM predicts that the answer to all eight questions should be yes. Since the counterfactual premise is treated as an intervention, the model infers that L and H are operating in each counterfactual scenario. The USM makes predictions that are fairly close to the predictions of the DMSM, but inferences about the deterministic separately caused system reveal one important difference. Because the rare cause L was observed to operate in the actual world, the DMSM assigns non-negligible probability to the conclusion that L is operating in the counterfactual scenario. The USM ignores all information about the actual world, and therefore generates a much lower probability. Of these two models, only the DMSM successfully predicts that the probability of L's operating is higher for the probabilistic separately caused system than for any other system.

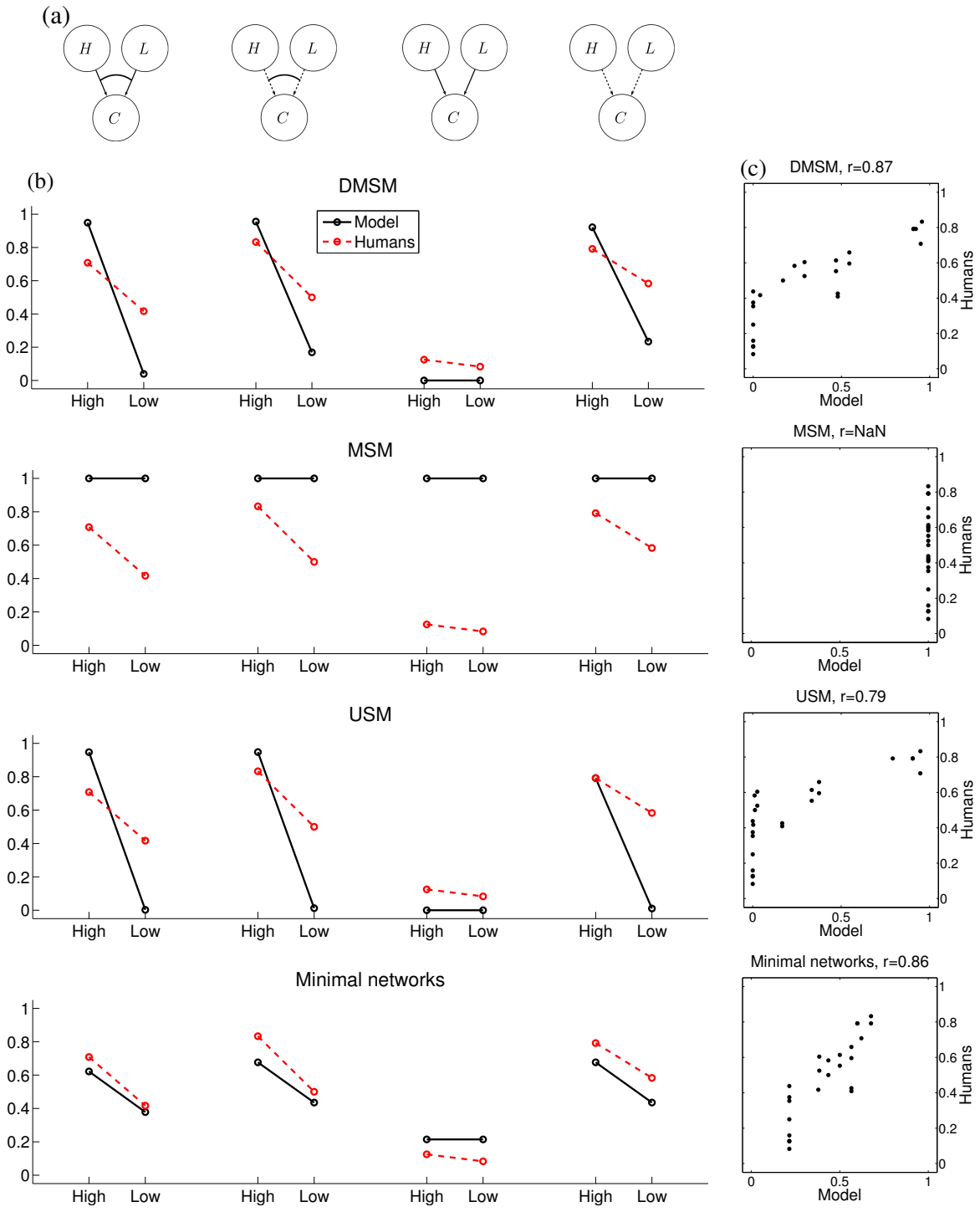

Figure 3: Model predictions and human responses for the experiments of Rips (2010). (a) Causal systems used in Rips's Experiment 3. Dotted lines represent probabilistic relationships and solid lines represent deterministic relationships. The edges linked by arcs represent AND relationships, where both causes must be active for the effect to occur, while un-linked edges represent OR relationships. Base rates differed between variables, with *H* or "High" variables occurring with probability .95 and *L* or "low" variables occurring with probability 0.05. (b) Results for counterfactual queries about the systems in (a). Proportions of "yes" judgments (human data) and probabilities (model predictions) are grouped according to the causal systems in (a). (c) Model predictions versus human judgments for all conditions across all four experiments in Rips (2010).

Because the MSM makes constant predictions across all of the different cases, an account in which individuals adopt either an MSM or USM strategy yields the same correlation with the data as the USM. Consequently, the DMSM fits the data better than accounts where some individuals always use $s = 0$ and others always use $s = 1$, or where each individual tosses a coin to set $s = 0$ or $s = 1$ on a question-by-question basis. Even so, our data do not decisively show that most individuals are characterized by intermediate values of s, and future empirical work is needed to address this question.

The final plot in Figure 3b shows predictions for the minimal network model described by Rips. Unlike the three models considered thus far, the minimal networks approach works with causal models such as Figure 1a instead of functional models such as Figure 1b. Given observations of the actual world (e.g. $L = H = C = 1$) and a counterfactual premise (e.g. $C = 0$), the minimal networks approach assumes that the counterfactual scenario contains a minimal set of breaks, or cases where variables differ from their actual values while their immediate causes do not.[2] In Hiddleston's (2005) original version of minimal networks theory, counterfactual queries receive affirmative answers only if they are true in *all* minimal networks. This version of the theory accounts poorly for the data in Figure 3, and the predictions shown there are based on a variant of the original theory that incorporates two additional assumptions suggested by Rips. First, if multiple configurations are minimal then participants are assumed to respond based on just one of these networks. With probability θ_1 participants sample one minimal network at random, and with probability $1 - \theta_1$ minimal networks are sampled according to their prior probabilities. Second, Rips proposes that with probability θ_2, a participant will ignore the evidence that they see and pick an answer at random. We set the parameters θ_1 and θ_2 to values that maximize the correlation between model predictions and human data.

Figure 3b shows that the minimal networks model accounts well for the data, and produces quantitative predictions that are superior to the DMSM. Note, however, that the DMSM has one free parameter and the minimal networks model has two. The DMSM can be adjusted in the same way as the minimal network model to allow for the fact that some participants responded randomly, and making this modification will bring the quantitative predictions of the DMSM into closer correspondence with human judgments.

We have focused so far on Experiment 3 of Rips (2010), but the scatterplots in Figure 3c summarize the performance of the four models across all four of Rips' experiments. The two best performing models are the DMSM and the minimal networks model. The DMSM therefore accounts for Rips' data as well as his own model despite requiring fewer free parameters. The DMSM performs better than the USM and the mixed-strategy account, but recall that the DMSM has one free parameter and the USM has no free parameters. Additional studies are therefore needed to confirm that sensitivity to observations about the actual world is critical when modeling human inferences about backtracking counterfactuals.

[2]Minimality of breaks is determined by set inclusion rather than counts—see Rips (2010) and Hiddleston (2005) for details.

Discussion

We presented a model of counterfactual reasoning that accounts for inferences about both counterfactual interventions and backtracking counterfactuals. Our approach is closely related to the modifiable structural model developed by Pearl and inherits the ability of this model to reason about counterfactual interventions. Our model, however, differs from the MSM in one critical respect: we allow for the fact that exogenous causal variables may take counterfactual values. We showed that this difference between the models allows the DMSM but not the MSM to account for Rips' experimental study of backtracking counterfactuals.

Although we believe that the DMSM is a step towards a unified theory of counterfactual reasoning, there are important theoretical and empirical questions that still need to be addressed. The DMSM accommodates both counterfactual observations and counterfactual interventions, but additional work is needed to characterize the conditions under which a generic counterfactual premise is interpreted as an observation or an intervention. A second direction for future work is to draw a sharper contrast between the DMSM and accounts that combine the predictions of the MSM and the USM using a weighted average, and to better understand individual differences in counterfactual reasoning.

Acknowledgments. This work was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative and by NSF award CDI-0835797.

References

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55-85.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.

Kemp, C., Shafto, P., & Tenenbaum, J. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1), 35–73.

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

Rips, L. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.

Sloman, S., & Lagnado, D. (2005). Do we 'do'? *Cognitive Science*, *29*(1), 5–39.