Back on track:
Backtracking in counterfactual reasoning

Tobias Gerstenberg (t.gerstenberg@ucl.ac.uk), **Christos Bechlivanidis** (c.bechlivanidis@ucl.ac.uk),
David A. Lagnado (d.lagnado@ucl.ac.uk)
Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP

Abstract

Would Dan have died if Bob hadn't shot? In this paper, we show that people's answer depends on whether or not they are asked about what would have caused Bob not to shoot. Something needs to change in order to turn an actual world into a counterfactual world. Previous findings of how people reason about counterfactuals have been mixed: sometimes people appear to backtrack and reevaluate the causes of a counterfactual state (e.g. Rips, 2010). At other times, people appear to treat counterfactuals like interventions that leave the past unchanged (Sloman & Lagnado, 2005). We experimentally manipulated the order in which participants were asked to consider the consequences of a counterfactual state. The results show that participants are more likely to backtrack when explicitly asked to consider a counterfactual's *causes*. However, when directly asked about the *effects* of a counterfactual state, most people don't backtrack.

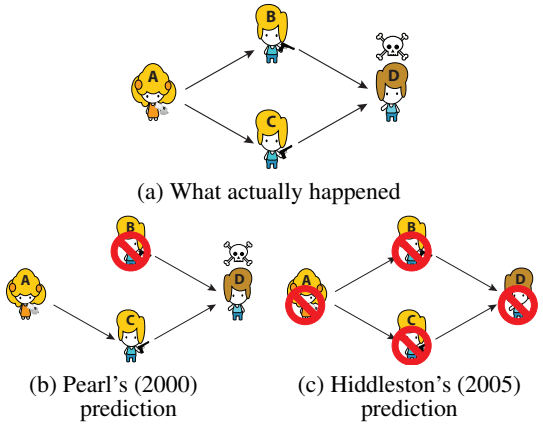
Keywords: counterfactuals; causality; inference; backtracking.

Introduction

Counterfactual thoughts play an important part in our everyday lives (see, e.g. Roese, 1997): if we had missed the submission deadline, you wouldn't be reading this paper. If we hadn't embarked on scientific careers, we would have become famous musicians. How do we evaluate the truth of such counterfactual statements? As life does not come with a rewind button, we can never know for sure. Hannes Kürmann, the protagonist in Max Frisch's play *Biography: A Game*, gets the unique chance to go back in time and play the game of life for a second time. However, despite full awareness of how his unhappy life will unfold and the firm belief that things could have turned out differently, Kürmann cannot bring himself to undo his past (and consequently, his present and future).

Max Frisch's play paints a rather fatalistic picture and suggests that counterfactual thoughts about how our life could have turned out differently are likely to be false. If everything happened as it actually did up until the point of the considered counterfactual, *it has to turn out false*. At some point, the counterfactual world has to diverge from the actual world in order to ensure the truth of the *if-part* (or antecedent) of a particular counterfactual statement. At least a change of mind would have been required to transform a scientist's life into that of a rock star.

Often there are a number of ways to realize the truth of a counterfactual's antecedent and the way in which we do so can sometimes have quite dramatic consequences. Consider the following situation: Anne is the commander of a firing squad and blows a whistle to signal to Bob and Chuck that it's time to shoot poor Dan (see Figure 1, cf. Pearl, 2000). Both Bob and Chuck shoot and Dan dies. Let us assume that



(a) What actually happened

(b) Pearl's (2000) prediction

(c) Hiddleston's (2005) prediction

Figure 1: If Bob had not shot, would Dan have survived?

the relevant causal relationships are deterministic: whenever Anne gives the signal, Bob and Chuck shoot and they never miss. Furthermore, each of Bob's and Chuck's shots are individually sufficient to bring about Dan's death. What do you think: would Dan have survived if Bob had not shot?

In this paper, we investigate how people evaluate counterfactual statements about simple devices that are structurally equivalent to the scenario just described. We first review theoretical frameworks that yield competing predictions about whether certain counterfactuals are true and then summarize previous empirical work on how people reason counterfactually. In a series of experiments, we test whether or not people spontaneously backtrack by manipulating the order in which participants are asked different counterfactual questions. We find that participants are more likely to backtrack when asked to explicitly consider the cause of the counterfactual's antecedent and suggest that the effect of question order can be explained in terms of a local processing strategy.

Theories of counterfactual conditionals

Let us illustrate the differences between theories of counterfactuals via the example of the counterfactual conditional "*If* Bob had not shot *then* Dan would have survived".

According to Lewis's (1979) account, the counterfactual conditional is true if the counterfactual world in which Bob had not shot ($B = 0$) and Dan would not have died ($D = 0$) is more similar to the actual world than any counterfactual world in which Bob had not shot ($B = 0$) but Dan would have died anyhow ($D = 1$). To generate the relevant counterfactual world, we are supposed to imagine a small miracle that transforms B from its original state to the considered counterfactual state and then let the counterfactual world unfold

according to the laws of nature. There are several problems with Lewis's account whereby most of which relate to the underspecified notion of similarity between different worlds (cf. Hiddleston, 2005). While Lewis aims to provide a non-causal account of counterfactuals and reduce causality to counterfactuals, others have argued that this puts the cart before the horse (Hiddleston, 2005; Pearl, 2000).

More recently, theories have been developed that take the notion of causality as primary and evaluate the truth of counterfactuals via reference to explicit causal assumptions that can be represented in causal Bayesian networks (CBN, Hiddleston, 2005; Pearl, 2000). In the spirit of Lewis's (1979) account, these theories evaluate the truth of counterfactuals by referring to similar worlds. However, they differ in how they conceptualize the causal similarity between different possible worlds.

According to Pearl's (2000) *pruning theory*[1], the evaluation of a counterfactual involves three steps. First, we update the values of the variables in the causal network based on our observations in the actual world (i.e. $A = 1$, $B = 1$, $C = 1$ and $D = 1$). Second, we change the value of the antecedent-variable (i.e. B) by means of an intervention. Such an intervention results in a mutilated causal network in which all incoming links to the intervened-on variable are removed (see Figure 1b). Third, we evaluate the consequent-variable (i.e. D) based on the variables' values in the mutilated network. Since the intervention in B disconnects all influences of upstream variables, A's value in the mutilated network remains unchanged. Because Chuck shoots whenever Anne gives the signal (i.e. $C = A$) and Dan dies if either Bob or Chuck shot (i.e. $D = max(B,C)$) the counterfactual is false. Dan would have died even if Bob had not shot.

Pearl's (2000) account of dealing with counterfactuals is similar to Lewis's (1979) in that the considered counterfactual world is identical to the actual world up until the point of the antecedent-variable. The antecedent-variable's counterfactual value is realized via an intervention that locally violates the causal relationships of the structure. The resulting counterfactual world is similar to the actual world in that the values of all variables that precede the antecedent-variable (or are causally independent from it) remain unchanged. However, it is dissimilar in that some of the causal relationships that were true about the actual world are not respected in the counterfactual world.

The opposite is true for Hiddleston's (2005) *minimal-network theory*. In this theory, the truth of a counterfactual conditional is evaluated by considering whether it holds in all worlds that are minimally different from the actual world but *consistent with its causal laws*. Given that the relationships between the actors in our scenario were described as deterministic, there are only two possible worlds that are causally consistent. The actual world (in which the values of all variables are 1) and a counterfactual world in which Anne did not give the signal, neither Bob nor Chuck shot and Dan survived (i.e. all values are 0). Hence, according to *minimal-network theory*, the considered counterfactual is true. If Bob had not shot, Dan would have survived (see Figure 1c).

The relevant counterfactual world is dissimilar from the actual world in that all events are different from how they actually were (including events that were temporally prior to the considered counterfactual). However, it is similar in that none of the actual causal relationships have been tampered with.

Note that evaluating the truth of counterfactuals according to *minimal-network theory* requires us to not only consider the consequences of the antecedent-variable. Bringing about the counterfactual state of the antecedent-variable in a way that is consistent with the causal laws requires us to backtrack and change the values of the antecedent-variable's *causes* as well. More generally, whereas *pruning theory* yields that backtracking counterfactuals (e.g. If Bob had not shot then Anne would not have given the signal) are always false, *minimal-network theory* holds that they can be true (at least in deterministic contexts).

Psychological studies of counterfactual reasoning

The results of previous studies on how people reason about counterfactuals have been mixed. Sloman and Lagnado (2005) found that people's counterfactual judgments are closely in line with the predictions of *pruning theory*. In one of their experiments, participants received descriptions of a causal structure identical to the one in the above scenario. In the abstract version of the task, they were informed that A causes B and C, and that B and C, in turn, each cause D. Knowing that D definitely occurred, participants answered the following two counterfactual questions: (a) If B had not occurred, would D still have occurred? (b) If B had not occurred, would A have occurred?

Pruning theory predicts that participants should answer 'yes' to both questions whereas *minimal-network theory* predicts negative responses. 80% of the participants answered 'yes' to (a) and 79% to (b). Responses were similar for scenarios in which the variables and causal relationship were described more concretely (a: 78%, b: 81%; averaged).

However, there has also been empirical support for *minimal-network theory* (Dehghani, Iliev, & Kaufmann, 2012; Rips, 2010; Rips & Edwards, in press). Rips (2010) and Dehghani et al. (2012) focused on backtracking counterfactuals and found that participants' judgments were sensitive to information about the base rates of the antecedent-variable's causes, the way in which these interact (disjunctive vs. conjunctive) and whether the causal links are deterministic or probabilistic. Since *pruning theory* rules out all backtracking counterfactuals, it cannot account for any of these effects. Recently, Lucas and Kemp (2012) have extended *pruning theory* to handle backtracking counterfactuals by allowing that variables which are not affected by the counterfactual intervention may take non-actual values.

One might argue that what answers a theory gives to backtracking counterfactuals is not of utmost importance for psychological theorizing. In everyday life, we are normally in-

[1] We follow Rips's (2010) terminology.

terested in the effects rather than the causes of counterfactuals. However, as the firing-squad scenario has shown, in some causal structures, whether or not a theory allows for backtracking also affects the truth of non-backtracking counterfactuals. Dan would have survived if Bob had not shot *only if* we backtrack and change Anne's action.

Rips and Edwards (in press) investigated participants' counterfactual reasoning using abstract devices that were structurally identical to the firing-squad scenario. They varied whether the causal links were described as deterministic or probabilistic (e.g. *A*'s operating *always*/*usually* causes *B* to operate) and whether *B* and *C* brought about *D* in a disjunctive ($D = max(B,C)$) or conjunctive manner ($D = min(B,C)$). Furthermore, they manipulated the framing of the counterfactual question between participants. Participants were either asked to consider that a certain component had *failed* or *not operated* (e.g. If *B* had not operated [failed] would *A*/*C*/*D* have operated?). Generally, participants tended to show less backtracking in the *failed* condition which suggests a local failure in the device than in the *not operated* condition. Furthermore, there was less backtracking for probabilistic compared to deterministic devices.

We will focus on structures with deterministic causal links for which the predictions between *pruning theory* and *minimal-network theory* dissociate strongest. Remember that for the deterministic disjunctive device, Sloman and Lagnado (2005) found that most participants answered positively to the question of whether *A* (or *D*) would have occurred if *B* had not occurred, Rips and Edwards (in press) found that in their *not operated* condition, almost all participants answered negatively. In the following, we will explore whether the way in which people mentally process counterfactual questions might account for these divergent findings.

Note that *pruning theory* and *minimal-network theory* make different predictions about what states of the system people need to consider when asked whether *D* would have operated if *B* had not operated. According to *minimal-network theory*, we first have to backtrack and infer that if *B* had not operated then *A* would not have operated. From this it follows that *C* and *D* would not have operated. *Pruning theory*, in contrast, predicts that we can evaluate the truth of the counterfactual *without* considering the state of *A* (see Figure 2). Since the counterfactual intervention on *B* does not affect the state of *C*, *D* is predicted to operate even if *B* had not operated (because of *C*).

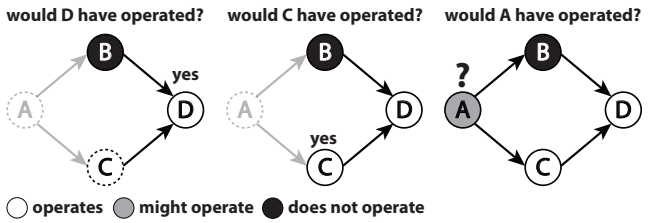


Figure 2: Hypothesized counterfactual reasoning process in the *D*–*C*–*A* condition.

This reasoning suggests that the order in which participants are asked to answer different counterfactual questions might influence how likely they show backtracking. In Sloman and Lagnado's (2005) experiment, participants were always asked about *D* first and then about *A*. In Rips and Edwards's (in press) experiment, participants were asked about *A*, *C*, and *D* and free to answer the questions in any order (cf. Figure 3a). Participants indicated their processing order on the response sheet and, generally, answered the questions from left to right (i.e. from *A* to *B*/*C* to *D*). In our experiments, we use a computerized task which allows us to manipulate the question order. Based on the discrepancy between Sloman and Lagnado's and Rips and Edwards's findings, we hypothesized that when asked to consider *A* before *D*, participants will be more likely to show backtracking than when asked about *D* before *A*. Note that neither *pruning theory* nor *minimal-network theory* predict any effects of question order.

Experiment 1: Replication

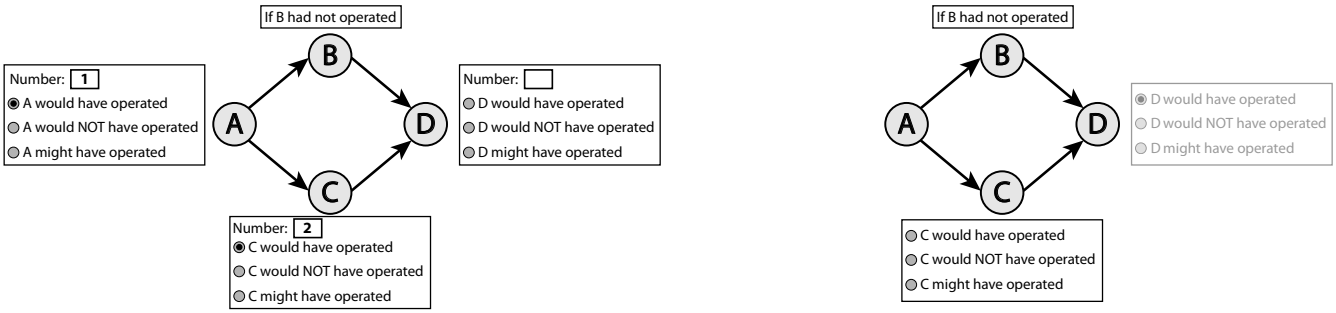
We first attempted to replicate Rips and Edwards's (in press) findings in the *not operated* condition using a computerized interface. Participants ($N = 40$, recruited via Amazon Mechanical Turk) saw eight different devices in randomized order and were asked to answer whether each of the other three components would have operated if *A*, *B* or *D* had not operated (i.e. 8 devices $\times$ 3 antecedent components $\times$ 3 consequent components = 72 questions).[2] The devices differed in whether the causal links were described as deterministic or probabilistic and whether *B* and *C* combined disjunctively or conjunctively. The probabilistic devices differed in whether (i) all links were probabilistic, (ii) only the links from *A* to *B* and *C* or (iii) from *B* and *C* to *D* (see Rips & Edwards, in press, for more details).

The order in which participants were asked about the different antecedent components was counterbalanced (*A*–*B*–*D* vs. *D*–*B*–*A*). For each antecedent component, participants were free to answer the counterfactual questions for the different consequent components in any order (see Figure 3a). For example, if *B* was the antecedent component (i.e. if *B* had not operated) a participant could answer about *A* (e.g. *A* would not have operated), *C* and *D* in any order. For each counterfactual, participants' response options were to say that the component *would have operated*, *would not have operated* or *might have operated*.

Results and Discussion

We followed Rips and Edwards's (in press) procedure and coded participants' responses as -1 (does not operate), 0 (might operate) and 1 (operates) in order to run standard statistical analyses. Figure 4 shows a selection of the results. Overall, we closely replicated Rips and Edward's findings with a correlation of $r = .92$ (*RMSE* = 0.23) between the

[2]Demos of the different experimental conditions can be accessed here: http://www.ucl.ac.uk/lagnado-lab/experiments/demos/backtracking_demo.html



(a) Experiment 1: Unconstrained question order

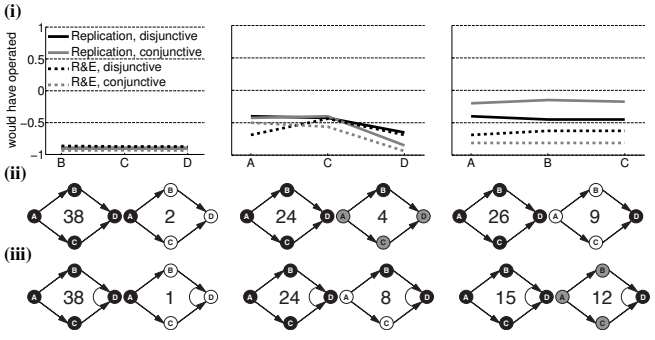
(b) Experiment 2: Constrained question order

Figure 3: Screenshots of the interface in Experiments 1 and 2 (*D–C–A* order condition).

averaged responses to the 72 questions in both experiments. Participants again tended to answer the counterfactual questions from left to right. For example, the average order in which participants indicated to have answered the counterfactual questions when *B* was the antecedent component was 1.23 for *A*, 2.10 for *C* and 2.67 for *D* (the corresponding values in Rips and Edwards's experiment were *A*: 1.44, *B*: 2.21 and *D*: 2.23).

Whereas both *pruning theory* and *minimal-network theory* predict the same pattern of responses when *A* is the antecedent component, their predictions differ when the antecedent components are *B* or *D*. *Minimal-network* predicts that the answers to all counterfactual questions are negative. *Pruning theory*, in contrast, predicts that when *D* is the antecedent, the answers to all consequent components should be positive. When *B* is the antecedent component, pruning theory predicts that the answers to both *A* and *C* should be positive. For the *D*, the answer is predicted to be negative for the conjunctive and positive for the disjunctive device.

In line with Rips and Edwards's findings and as predicted by *minimal-network theory*, a majority of participants answered the counterfactual questions negatively. For example, when asked whether *A*, *C* and *D* would have operated if *B* had not operated for the conjunctive device, 24 participants showed backtracking whereas only 8 participants responded in line with *pruning theory* (see Figure 4b).



(a) If *A* had not operated

(b) If *B* had not operated

(c) If *D* had not operated

Figure 4: (i) Mean judgments separated for the disjunctive and conjunctive deterministic device. The labels on the x-axis correspond to the consequent-components. Most frequently endorsed structures for the (ii) disjunctive and (iii) conjunctive devices. *Note*: R&E = Rips and Edward's (in press) data.

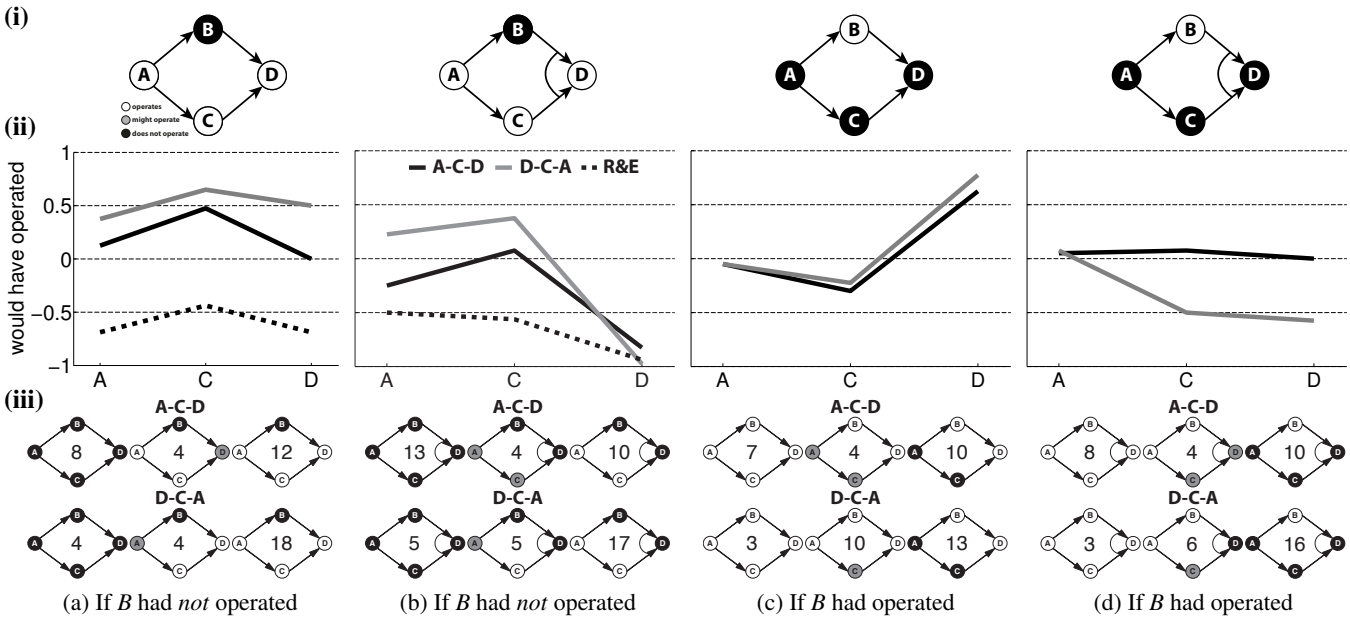
Experiment 2: Order Manipulation

Having replicated Rips and Edwards's finding, Experiment 2 tests the hypothesis that the order in which participants are asked to answer different counterfactual questions influences the degree to which they backtrack. With *B* as the counterfactual antecedent, we predicted that participants will show more backtracking when asked about *A* before *D* and less backtracking when asked about *D* before *A*.

Between participants ($N = 320$, recruited via Amazon Mechanical Turk), we manipulated the question order (*A–C–D* vs. *D–C–A*), whether the device was disjunctive or conjunctive as well as whether, in actuality, all or none of the components were operating (40 participants per condition). When all components were operating participants were asked to consider the counterfactual that *B* had *not* operated (see Figure 5a and b). When none of the components were operating, participants considered that *B* had operated (see Figure 5c and d).

Our processing hypothesis predicts an interaction between the question order, the type of device and its actual state. The question order is predicted to influence participants' judgments about the counterfactual state of *D* for (a) the disjunctive device in which everything is actually operating and (d) the conjunctive device in which nothing is operating (see Figure 5a and d). In these cases, whether *D* would have been different from actuality depends on whether or not participants backtrack. Accordingly, we predicted that participants in the *D–C–A* condition are more likely than participants in the *A–C–D* condition to say that *D* would have operated for (a) and less likely to say that *D* would have operated for (d). In contrast, we do not predict an effect of question order for devices (b) and (c). The counterfactual state of *B* is by itself sufficient to bring about a change in *D* without the need to consider the states of the other components.

Figure 3b shows a screenshot of the *D–C–A* condition. Participants first only saw the text box for *D*. Having answered that question, the response was locked and the next text box



(a) If *B* had *not* operated (b) If *B* had *not* operated (c) If *B* had operated (d) If *B* had operated

Figure 5: Mean judgments (ii) and frequency of endorsed networks (iii) for different causal devices (i). *Note*: For each device, the leftmost networks in (iii) are predicted by *minimal-network theory* and the rightmost networks by *pruning theory*. The networks in the middle are the most frequently endorsed networks predicted by neither of the two theories.

appeared. All participants just provided answers for a single device.

Results and Discussion

For ease of interpretation, we analyze the results for devices in which everything is operating initially (Figures 5a and b) and in which nothing is operating (Figures 5c and d) separately and focus on participants' answers to component *D*.

For the operating devices (a, b), there was a significant main effect of structure, $F(1,156) = 44.13, p < .001, \eta_p{}^2 = .459$ and no main effect of question order ($p = .097$). Participants were more likely to think that *D* would have operated for the disjunctive (*M* = 0.25, *SD* = 0.88) compared to the conjunctive device (*M* = -0.9, *SD* = 0.41).

More interestingly, there was a significant interaction between structure and question order $F(1,156) = 9.59, p = .002, \eta_p{}^2 = .058$. For the disjunctive device, participants in the *D–C–A* condition were more likely to say that *D* would have operated (*M* = 0.5, *SD* = 0.82) than participants in the *A–C–D* condition (*M* = 0, *SD* = 0.88), $t(78) = -2.64, p = .01, d = -0.6$. For the conjunctive device, there was no significant difference as a function of question order ($p = .101$).

The results for the non-operating devices (c, d), closely mirrored the results of the operating devices. Again, there was a significant effect of structure, $F(1,156) = 39, p < .001, \eta_p{}^2 = .302$ and no main effect of question order ($p = .079$). Participants were more likely to think that *D* would have operated for the disjunctive (*M* = 0.70, *SD* = 0.68) compared to the conjunctive devices (*M* = -0.29, *SD* = 0.87).

The interaction between structure and question order was significant $F(1,156) = 5.26, p = .003, \eta_p{}^2 = .055$. While there was no significant difference of question order for the disjunctive device ($p = .329$), in the case of the conjunctive device, participants in the *D–C–A* condition were less likely to say that *D* would have operated (*M* = -0.58, *SD* = 0.78) than participants in the *A–C–D* condition (*M* = 0, *SD* = 0.88), $t(78) = 3.1, p = .003, d = 0.70$.

These results demonstrate that the order in which participants were asked about the different components affected whether they believed that *D* would have operated. In the *A–C–D* condition, 36 participants (out of 160) answered as predicted by *minimal-network theory* and 42 as predicted by *pruning theory* (see Figure 5iii). These numbers shifted towards much less backtracking in the *D–C–A* condition: only 15 participants answered consistently with *minimal-network theory*, whereas 68 answered in line with *pruning theory*.

The results also revealed another interesting pattern: the absolute value of participants' averaged answers about component *A* (*M* = 0.15) were generally less certain (i.e. closer to 0) than their answers about *C* (*M* = 0.33) and *D* (*M* = 0.53). The shift towards averaged 0 responses from component *C* to *A* in Figures 5a and d for the *D–C–A* condition is neither predicted by *pruning theory* nor *minimal-network theory*. We consider this to be evidence that people process counterfactual questions in a more local fashion rather than simultaneously considering the states of all variables in the system.

For example, when asked whether *D* would have operated if *B* had not operated (cf. Figure 5a) most participants in the disjunctive *D–C–A* condition answer 'yes' to *D*. Having answered positively to *D* commits participants to saying that *C* would have operated as well (cf. Figure 2). Otherwise, there is no explanation for why *D* operates. However, when considering *A*, participants have reached a state of causal inconsistency. Having answered 'yes' to *C* but knowing that *B* did not

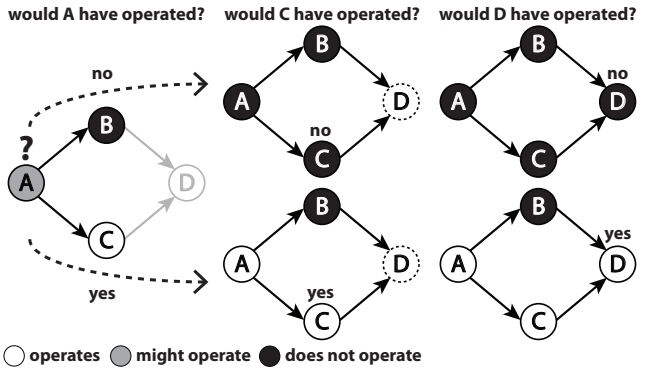


Figure 6: Hypothesized local counterfactual reasoning process in the *A–C–D* condition.

operate, they can either resolve this inconsistency by answering 'yes' to *A* and assuming a fault in *B*. Alternatively, they can answer 'no' to *A* and assume that *C* must have operated spontaneously. The same rationale also explains the pattern of results for device (d). For devices (b) and (c), participants' response to *D* does not commit them to a particular response for component *C* — the counterfactual state of *B* already accounts for the change in *D*.

Participants in the *A–C–D* condition have to resolve the potential causal inconsistency right at the start (see Figure 6). As the results show, participants are split in how they do so: some backtrack and respond in line with *minimal-network theory*. Others don't and respond as predicted by *pruning theory*.

General Discussion

The capability to think about possible states of the world and reason through what would or could have happened is one of the hallmarks of human cognition. Counterfactual thoughts are of central importance to attributions of responsibility (Lagnado, Gerstenberg, & Zultan, accepted) and causality (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012). In this paper, the aim was to gain insight into people's counterfactual processing. Based on mixed findings in previous research (Dehghani et al., 2012; Meder, Hagmayer, & Waldmann, 2009; Rips, 2010; Sloman & Lagnado, 2005), we investigated whether the order in which participants are asked to reason about the consequences of certain counterfactual states could shed light on these inconsistencies. We first replicated Rips and Edwards's (in press) experiment and then manipulated the order of counterfactual questions in an identical experimental setup.

As hypothesized, participants' answers were more in line with the predictions of *minimal-network theory* (Hiddleston, 2005) when asked to consider a possible *cause* of the counterfactual state first. In contrast, when participants were asked to consider the *effect* of a counterfactual state first, participants showed less backtracking and followed the predictions of *pruning theory* (Pearl, 2000) more closely. However, the overall pattern of results was not predicted by either theory. We discussed that a more local processing strategy is consistent with this data (cf. Fernbach & Sloman, 2009, for a similar idea in causal learning). Accordingly, when asked to consider a certain counterfactual, people do not spontaneously think through the implications that this counterfactual state has for the whole system. Rather, participants' responses are indicative of a more local processing strategy that considers only parts of the system. The order in which participants are probed about the counterfactual world hence has a significant effect on what changes they make in order to account for the stipulated counterfactual state. Applied to our initial example, whether Dan is believed to have survived if Bob had not shot depends on whether we are asked to consider Anne first.

While the results of Experiment 1 have shown that participants' responses were closely in line with *minimal-network theory*, the results in Experiment 2 were more mixed. In future research, we aim to (i) generalize these findings using less abstract stimuli and (ii) investigate more closely what differences between the reported experiments account for participants' tendency to backtrack or not. We speculate that both the explicit contrast between deterministic and probabilistic systems as well as the fact that participants have to think through a great number of different devices, encourages them to endorse a more holistic strategy that favors responses that are causally consistent. However, when not asked explicitly to consider the causes of a counterfactual state, most participants stay on track and don't backtrack.

Acknowledgments

We thank Lance Rips for providing the data and Brian Edwards for insightful comments. This work was supported by a doctoral grant from the AXA research fund (TG) and an ESRC grant RES-062-33-0004 (DL).

References

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

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

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

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

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (accepted). Causal responsibility and counterfactuals. *Cognitive Science*.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lucas, C. G., & Kemp, C. (2012). A unified theory of counterfactuals. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249–264.

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

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

Rips, L. J., & Edwards, B. (in press). Inference and explanation in counterfactual reasoning. *Cognitive Science*.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–148.

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