UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Backwards counterfactuals and the closest possible world

Permalink

https://escholarship.org/uc/item/0q93p7sc

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Grosu, Ioana

Le, Dominic

Ganea, Patricia

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Backwards counterfactuals and the closest possible world

Ioana Grosu (ioana.grosu@utoronto.ca)
Department of Applied Psychology & Human Development, University of Toronto

Dominic Le (dominic.le@mail.utoronto.ca)
Department of Applied Psychology & Human Development, University of Toronto

Patricia Ganea (patricia.ganea@utoronto.ca)
Department of Applied Psychology & Human Development, University of Toronto

Abstract

One source of complexity in counterfactual reasoning is the order in which events are presented within the conditional. Counterfactuals with a backwards order of events (aka 'backtracking' counterfactuals) involve reasoning backward: from the consequent to the antecedent. We extend on prior experimental work (e.g., Rips 2010), and consider the possibilities adults reason over when they backtrack. We find that adults' reasoning strategies tend to be inconsistent when responding to backtracking questions. In scenarios involving a single causal variable, participants do not generally allow for extraneous changes from the actual world. Furthermore, when reasoning forward along a causal chain, participants do not allow for extraneous changes. However, in backtracking scenarios involving multiple causal variables, participants are at chance in choosing worlds with extraneous changes. We provide novel evidence for the changes allowed from the actual world when backtracking, with mixed support for theoretical claims such as Minimal Networks Theory.

Keywords: possibility reasoning, semantics, counterfactuals, causality, backtracking

Introduction

Counterfactual (CF) reasoning allows us to consider how the world might have been, given a change to our actual state of affairs. It requires an ability to reason over non-actual events, and is tied closely to our understanding of causality (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; German & Nichols, 2003; Henne, Kulesza, Perez, & Houcek, 2021; Mackie, 1980; Spellman & Mandel, 1999; Mandel, 2003). A counterfactual conditional (e.g., "If dogs had wings, they would fly") is comprised of an antecedent ("If dogs had wings") and a consequent ("They would fly"). While the antecedent typically describes an event that happens before the consequent, there exists a special class of counterfactuals in which this is not the case. These are referred to as *backtracking*[1] counterfactuals, since they involve reasoning backwards from the consequent to the antecedent (Lewis, 1979; Frisch, 2005; Hiddleston, 2005; Ward, 2014; Khoo, 2016).

To illustrate, (1) is a *backtracking counterfactual.*

(1) If the dog had barked, he would have seen the mailman.

Here, we reason backwards from the *consequent* event (the dog seeing the mailman) to the *antecedent* event (the dog barking), since the consequent event occurs before the antecedent event. Compare this to the *forward counterfactual* in (2):

(2) If the dog had seen the mailman, he would have barked.

Here, (contra (1)) one reasons forward when analyzing the counterfactual scenario, from the antecedent event (the dog seeing the mailman) to the consequent event (the dog barking).

While forward counterfactuals are well-attested in the literature on adult reasoning ability, backwards counterfactuals are less commonly discussed and relatively understudied in experimental work. However, there is a growing body of research showing that adults reason backwards when forwards reasoning is unlikely (Rips, 2010; Rips & Edwards, 2013; Gerstenberg, Bechlivanidis, & Lagnado, 2013; Han, Jimenez-Leal, & Sloman, 2014).

A complete theory of counterfactuals must not only accurately account for forward counterfactuals, but also for backwards counterfactuals (Lassiter, 2018). One popular account is Minimal Networks Theory (MNT) presented in Hiddleston (2005). However, this account has demonstrated limitations in the backtracking cases for which it can account (Rips, 2010; Rips & Edwards, 2013).

In the present study, we expand on prior work analyzing backtracking within the context of minimal networks, by studying the nature of the changes participants allow to a causal network. We consider different types of causal structures (with both one and multiple causes which can lead to some effect). Working within the framework of causal networks to design our stimuli, we consider whether participants allow for extraneous changes to a causal network (e.g., changing a variable not mentioned in the antecedent, which is not causally downstream from the effect under consideration).

To this end, we address the following question: *When adults reason backwards over possibilities, do they tend to take a parsimonious approach to changes in the causal system?* That is, do they prefer to keep constant variables which are unmentioned in the antecedent? For forward counterfactuals, this is suggested to be the case (Lewis, 1973), but the counterfactual changes we allow when backtracking remains an open question.

[1] or *backward counterfactuals* in Khoo (2016)

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

We conduct an experimental task on adults, in which we manipulate the nature of the causal network, and require participants to respond to both back and forward-tracking counterfactual questions. Unlike prior studies, we prompt for a minimal (vs. non-minimal) network in our questions, to determine whether participants truly have a preference for minimal alterations to a causal network. We find that in scenarios where the unmentioned variable is not a cause of the antecedent, participants choose to leave it unchanged. However, when the unmentioned variable is a cause of the antecedent, participants do not consistently choose the minimally altered network. This builds upon prior work such as Rips (2010), showing that participants not only are not consistently following the response patterns expected for MNT, but that they choose a non-minimally altered possibility to a similar degree as the minimally-altered possibility.

Background

Lewis' account

Backtracking counterfactuals add a layer of complexity for theoretical accounts of counterfactuality, especially for Lewis (1979). The existence of backtracking counterfactuals is contradictory to the asymmetric view in Lewis - i.e., that the past does not depend on the present (since backtrackers are an example of such a dependence). Lewis avoids the issue by claiming that backtracking counterfactuals require a *special resolution of vagueness*, in which the past can change depending on the present state, and excluding backtrackers from the standard analysis of counterfactuality. The Minimal Networks (MNT) approach in Hiddleston (2005), on the other hand, provides a causal models-based account which allows for both forward and backtracking.

In order to understand MNT we must first appeal to a notion of similarity when evaluating counterfactuals (Lewis, 1973). When we reason over counterfactuals, we have fairly sharp judgments about the truth or falsity of the utterances - which is not possible under standard truth-conditional semantics, since the antecedent of a counterfactual is, by its nature, false.

Here, we present a simplified version of Lewis's similarity principle, which we build on for our discussion of MNT. We refer to this as the *closest possible world constraint* (CPWC).

> *A counterfactual A (antecedent) therefore C (consequent) is true if C is true in all worlds in which A is true and that are otherwise most similar to the actual world.*

If we take our initial example ("If dogs had wings, they would fly"), and apply this constraint, we would first have to consider worlds in which the antecedent is true. That is, the worlds in which dogs have wings. Of these worlds, the worlds in which dogs would fly are closer to our state of affairs than worlds in which, e.g., dogs would become herbivores.

Building on the CPWC, MNT allows us to formalize the exact nature of the changes to the world, and consider the models of the world which we generate when reasoning counterfactually. It also provides a potential account for backtracking, though experimental evidence shows that it does not generate accurate predictions for all types of causal systems.

Causal networks and counterfactual reasoning

Accounts of counterfactuality such as Hiddleston (2005) and Pearl (2009) rely on causal networks to conceptualize counterfactual reasoning. Causal networks are variations of directed acyclic graphs meant to map out causal relationships (Danks, 2014). These graphs are comprised of nodes and links between nodes. The nodes comprise the set of variables, and the links between nodes represent the relationships between the variables. The utility of using causal networks (and more broadly, causality) to account for counterfactual and conditional reasoning is highlighted in, e.g., Pearl (2009), Spirtes et al. (2000) and Tversky and Kahneman (1980), and the validity of causal networks as a way of modeling causal reasoning is discussed in Gopnik et al. (2004), where it was found that inferences made by children and infants with respect to causal scenarios are consistent with inferences resulting from causal networks.

Following Hiddleston (2005), we define similarity based on the changes one makes to the causal structure representing the actual world.

To illustrate, consider the following scenario:

> *Amy needs to get to school on time. In order for her to get to school, two things need to happen. She needs to set her alarm the night before, and there can't be a lot of traffic on the roads. Today, the roads are clear, and she sets her alarm. Therefore, she gets to school on time.*

Here, the set of facts is the following:

R: *Roads are clear*
A: *Alarm is set*
S: *Amy gets to school on time*

And the law governing these facts is restated below:
If the roads are clear (R=1) and her alarm is set (A=1), Amy gets to school on time (S=1).

The causal network for the sample scenario is in Figure 1.

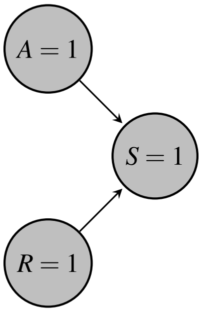
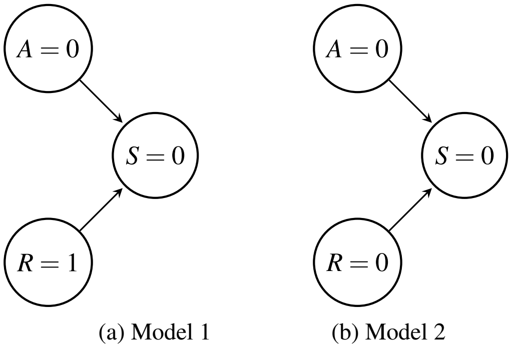


Figure 1: Sample scenario causal diagram

Now suppose that Amy did not set her alarm on a certain day (i.e., $A = 0$ is our antecedent). From this antecedent, we can generate two possible causal diagrams, which are shown in Figure 2.



(a) Model 1 (b) Model 2

Figure 2: Causal diagrams given $A = 0$

Each scenario is summarized below:

Model 1: If Amy had not set her alarm ($A = 0$), the roads would still have been clear ($R = 1$), but she would not have gotten to school on time ($S = 0$).
Model 2: If Amy had not set her alarm ($A = 0$), the roads would still have not been clear ($R = 0$), and she would not have gotten to school on time ($S = 0$).

Similarity is defined based on the number of causally independent facts which are preserved from the actual world. In order for a fact to be causally independent it must not be caused by the antecedent fact (i.e., it is not directly downstream from the antecedent). The model which maintains the most causally independent facts is the minimally altered model (i.e., the model with the minimal number of alterations to it with respect to its variables). In this case, Model 1 maintains more causally independent facts ($R = 1$ is not changed) than Model 2. Model 2 changes the other unmentioned fact ($R = 0$) as well. One would therefore predict the scenario in Model 1 to be judged as more plausible, which is in line with intuitive judgments of the two scenarios.

Prior experimental studies

There is a growing body of experimental studies which have considered the way in which theoretical accounts of backtracking can align with observed participant data (Dehghani, Iliev, & Kaufmann, 2012; Sloman & Lagnado, 2005; Han et al., 2014; Gerstenberg et al., 2013; Rips, 2010; Rips & Edwards, 2013; Lucas & Kemp, 2012, 2015). Typically, these accounts either use a minimal-networks type approach (e.g., Dehghani et al. (2012) and Rips (2010)) or a Pearl-style approach (e.g., Lucas and Kemp (2012)'s DMSM theory). While we focus on a Hiddleston-style notion of similarity here as the basis for our analysis, it is important to note some of the factors which can affect participants' ability to backtrack.

For instance, Han et al. (2014) found that participants are less inclined to backtrack if it is unnecessary for reaching a conclusion —when another route of reasoning such as forward-tracking suffices. The order of questions also can affect the selection of causal structures, as seen in Gerstenberg et al. (2013). Additionally, prior knowledge can also affect the way in which participants respond to backtracking prompts, as participants may be susceptible to certain cognitive biases when reasoning counterfactually (Dehghani et al., 2012). From prior experimental work, however, it does become clear that participants do not exclusively engage in forward-tracking when faced with counterfactual prompts, and they take into account prior variables and can change them as needed.

However, the nature of these changes still remains largely understudied. In particular, the degree to which participants are flexible in allowing changes to causally upstream variables, and whether an account such as MNT can be predictive of such changes. In this paper we mostly build off of experimental work in Rips (2010) (and Rips and Edwards (2013), which follows up with further experimental manipulations). In Rips (2010), the predictions of MNT were tested in a series of four experiments (three with backtracking and one with forward). The experiments vary in their use of different causal structures and phrasing, as well as whether the connections between variables are deterministic or probabilistic. The test questions prompt for the status of a variable as either a question (of the sort "If component X had not operated would component Y have operated?"), or as a binary choice ("If component X were not operating, component Y would be operating") where participants have to decide if the sentence "follows" or does not "follow".

Take the device in Figure 3 as an example, from Rips (2010)'s Experiment 1.

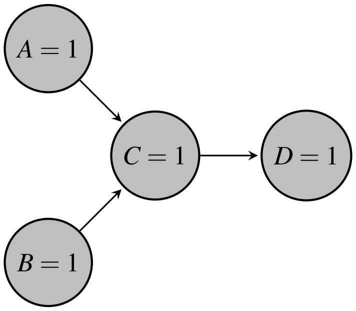


Figure 3: Rips (2010), deterministic jointly caused device

Here, the rule for the device is that it is jointly caused. This means that *both* A and B must have their values set to 1 in order for C (and therefore D) to also be 1. Participants would be asked, in this case "If Component C had not operated, would component A have operated?". In this case, the MNT prediction would be "no", since there are two plausible minimal networks ($A = 1$ and $B = 0$ or $A = 0, B = 1$), and A is not 1 in one of the networks. However, for this test item, around half of the participants responded with 'yes'.

While the results show that variables upstream from the antecedent can have an impact on counterfactual responses,

responses across test items tended to be fairly evenly split, therefore not providing conclusive evidence for a minimal networks approach. Crucially, none of the conditions presented the entirety of the potential counterfactual models to participants. Since the status of the unmentioned variable in the causal systems is not evident, one cannot accurately gauge whether participants have a preference for minimally altered worlds, or whether they are more flexible in the changes they allow to a causal system.

For example, in the above test item, it is possible that the participants responding with 'no' are considering a non-minimally altered world instead (one in which both A and B are set to 0), or that they are simply employing some form of basic conditional reasoning, and only considering at the causal connection between A and C, without considering the status of B. In addition, some of the 'yes' responses could be due to participants changing only the status of the B, and opting to maintain A. The purpose of current study is to address some of these open questions, and provide a more targeted set of possibilities for participants to choose from. In doing so, we aim to better understand the models which we reason over when backtracking over simple causal systems.

Methods

Participants

We tested 46 native English-speaking adults from the United States. Participants signed up to our study via Amazon Mechanical Turk. Upon signing up, participants were provided with an external link to an online Qualtrics survey. Participants were given one hour to complete the survey at their own pace, unsupervised. We analyzed responses from 36 participants, excluding 10 for incorrectly answering at least one control question (out of 3 per scenario). Although our exclusion criteria were strict, we found that participants largely understood the control questions: by scenario, control questions were correctly answered at a rate of 94% (WITH-INERT), 92.75% (DISJUNCTION), and 97.83% (CONJUNCTION).

Materials and Procedure

In this study, we aim to closely analyze the changes that participants allow from the actual world, and whether these changes are parsimonious. To do this, we present participants with three different causal structures, varied in terms of their rules and their complexity. Once we ensure that participants are able to identify the causal structure through a series of control questions, we prompt them with counterfactual changes to the world (either a change to one of the causes (*forward-tracking*), or a change to the effect (*backtracking*)). Participants are asked to choose between two possible worlds: either (1) a minimally different world and (2) non-minimally altered world. While both choices are viable (i.e., are antecedent-worlds and follow the law of the system), we would expect participants who use a parsimonious approach to counterfactual changes to not choose world (2) in favor of (1).

Expanding on previous experimental work on backtracking (e.g., Rips (2010)), we do not present participants with yes/no questions about the occurrence of a counterfactual event. Instead, we ask participants to choose between two worlds, in order to have a clearer idea of their reasoning strategy, and their preference over models. For multi-cause backwards CF reasoning, findings from prior studies' results cannot discriminate usage of the CPWC (since there is no clear prompting for a non-minimally altered world). Assuming that participants understand how the states of each variable influences the state of other variables, our research question targets whether the CPWC applies in both temporal directions during counterfactual reasoning about causal systems.

Scenarios We presented participants with short stories, all of which were audio recorded by the second author (via Adobe Premiere Pro) with corresponding visuals (created by the second author via Canva). Scenarios involved simple machines, with two blocks which could be placed on a box and make the box light up, based on different rules. The machines were based on the *blicket detector* paradigm used in Gopnik and Sobel (2000), Gopnik, Sobel, Schulz, and Glymour (2001), Nyhout and Ganea (2019), i.a.

For our test questions we manipulated two factors. The first was the causal direction of the clauses in the prompt (FORWARD, BACKWARD). The second was the causal structure of the scenario (WITH-INERT, DISJUNCTION, CONJUNCTION). DISJUNCTION and CONJUNCTION cases were qualitatively organized as multi-cause systems in comparison to the single-cause WITH-INERT system.

For each system there exist at least two possible worlds of varying closeness to the actual world. Each system has a distinct causal mechanism, and the systems together encompass a set of basic logical rules. All systems involve three variables - a light, and two colored blocks which differ in whether they can turn the light on by themselves, whether they need another block, or whether they do not turn the light on at all.

In our single-cause (WITH-INERT) case, the blue block can turn the light on by itself, and the orange block does not turn on the light. Participants are presented with the verbal instructions, along with the accompanying pictures.

Our other two (multi-cause) scenarios follow a similar structure in their presentation. One of the systems is CONJUNCTION (or jointly caused), where both blocks (purple and green) are required to be on the lightbox in order for the light to turn on. The other system is DISJUNCTION (or separately caused), and either one of the blocks (yellow or pink) can turn on the light by itself.

Control questions After presenting participants with the basic setup of the causal structure, we asked them a series of control questions to ensure that they understood the scenarios. We excluded participants from our sample if they could not accurately respond to these questions.

Using our WITH-INERT scenario as an example, we presented participants with the following questions, and accompanying images. All control questions were presented as short videos, with a binary-choice yes/no response after.

- “Now, the blue block goes on the box, and the orange block stays off the box. Will the light turn on?” (yes/no)
- “Now, the orange block goes on the box, and the blue block stays off the box. Will the light turn on?” (yes/no)
- “Now, the blue block goes on the box, and the orange block goes on the box. Will the light turn on?” (yes/no)

A similar set of control questions was used for the multi-cause conditions as well, in the same format.

Test questions Figure 4 outlines the counterfactual test questions across conditions.

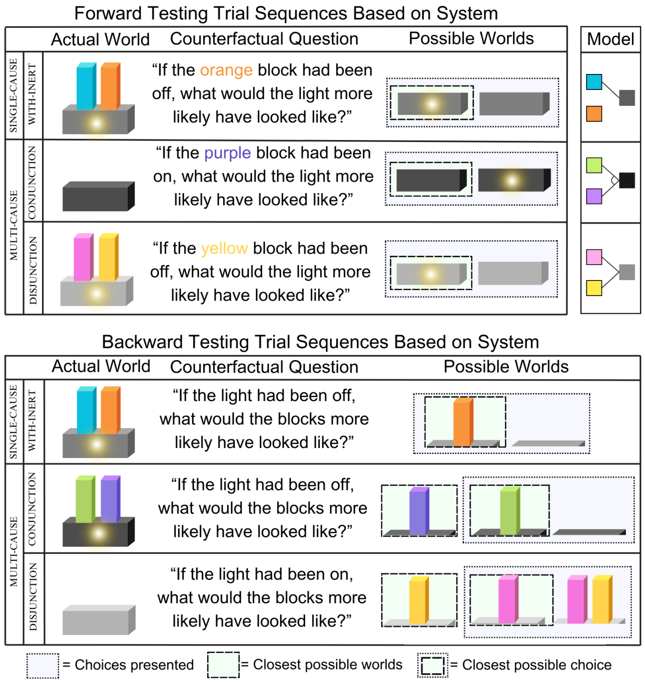

Forward Testing Trial Sequences Based on System
Actual World
Counterfactual Question
Possible Worlds
Model
SINGLE-CAUSE
WITH-INERT
“If the orange block had been off, what would the light more likely have looked like?”
MULTI-CAUSE
CONJUNCTION
“If the purple block had been on, what would the light more likely have looked like?”
DISJUNCTION
“If the yellow block had been off, what would the light more likely have looked like?”
Backward Testing Trial Sequences Based on System
Actual World
Counterfactual Question
Possible Worlds
SINGLE-CAUSE
WITH-INERT
“If the light had been off, what would the blocks more likely have looked like?”
MULTI-CAUSE
CONJUNCTION
“If the light had been off, what would the blocks more likely have looked like?”
DISJUNCTION
“If the light had been on, what would the blocks more likely have looked like?”
= Choices presented
= Closest possible worlds
= Closest possible choice

Figure 4: Test conditions, with counterfactual prompts and possible worlds (blocks that were not on the box were shown beside the box in the materials presented to participants and not shown here to save space).

Backwards and forward counterfactual reasoning was tested across the three causal structures. For each condition, participants were presented with two possible worlds - one of which is minimally altered (i.e., no unnecessary changes were made in order to ensure the truth of the antecedent), and the other which is not. Participants were asked to choose the most likely option between the two worlds, as a binary forced-choice task. For example, in BACKWARD CONJUNCTION, the actual world involves both blocks on the box. Counterfactually, participants were told that the light is off. Then, they chose between a world in which the green block is still on the box (minimally altered) or a world in which no block is on the box. The actual world and counterfactual questions were presented as short videos to participants, and participants were asked to click on the image corresponding to the world state they found most likely, given the antecedent.

Procedure Each participant saw all three scenarios (WITH-INERT, CONJUNCTION, DISJUNCTION) in randomized order. For each scenario, they are presented first with a video of the lightbox setup and causal connections. After we introduced each box, we asked participants the associated set of control questions. The order of control questions was randomized between trials. Responses to each control question was collected, and we did not correct participants‘ answers. Following this, we asked participants both a BACKWARD counterfactual question and a FORWARD counterfactual question in random order. For each of our test questions we coded each participant‘s response score as 1 if it was the closest possible world and 0 otherwise.

Results

From a preliminary analysis, participants preferred the closest possible world (i.e., the world with the fewest non-causally downstream changes) across FORWARD tracking conditions (96.3%) as would be predicted by Hiddleston (2005) while they preferred the closest possible world less so for the BACKWARD conditions (56.5%), contra Hiddleston. As shown in Figure 5, responses across the FORWARD condition were consistently close to ceiling, while responses across the BACKWARD conditions were mixed. Specifically, responses in the SINGLE-CAUSE conditions indicated a preference for the closest possible world, versus in the MULTI-CAUSE conditions, where participants were consistently close to chance in their selection of worlds.

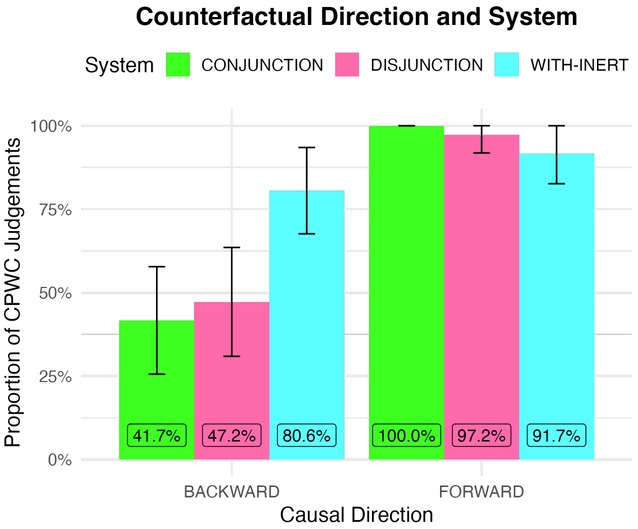


Figure 5: Percentage of responses that align with the Closest-Possible-World Constraint (CPWC) across tracking directions and systems with Standard Error Bars

Given the repeated-measures (within-subject) nature of the data, we fitted a Generalized Estimating Equation model via the geepack package in R. The dependent variable was the participant's score, which was assumed to be discretely binomially distributed between 0 and 1. The independent variables were the scenario (WITH-INERT, CONJUNCTION, DISJUNCTION) and the causal direction (BACKWARD, FORWARD).

Participants generally made more CPWC judgments in the FORWARD than the BACKWARD CF direction, ($b = 3.138, SE = 0.665, \text{Wald} = 22.27, p < .001$). Interestingly, participants preferred the closest possible world more in WITH-INERT than CONJUNCTION, $b = 1.180, SE = 0.452, \text{Wald} = 6.82, p < 0.01$, with no significant difference between the MULTI-CAUSE conditions, $b = 0.095, SE = 0.372, \text{Wald} = 0.07, p = .747$, across CF directions.

A series of binomial tests was conducted to test whether selection of the closest possible world was at chance (p = 0.5). Significance was evaluated using Bonferroni corrections for multiple tests. BACKWARD MULTI-CAUSE responses did not differ from chance (CONJUNCTION: p_{adjusted} = 1.00, 95% *CI* [0.255, 0.592]; DISJUNCTION: p_{adjusted} = 1.00, 95% *CI* [0.304, 0.645]). However, the BACKWARD WITH-INERT condition was significantly greater than chance, $p_{\text{adjusted}} < .001$, 95% *CI* [0.640, 0.918]. All FORWARD conditions as well were greater than chance ($p_{\text{adjusted}} < .001$).

Discussion

We observe that participants did not choose the closest possible world in either of the multi-cause backtracking conditions. However, in the SINGLE-CAUSE condition, participants consistently preferred the closest possible world whether they reasoned forwards or backwards along a causal chain of events. Additionally, in both of the MULTI-CAUSE FORWARD conditions, participants were at (or near) ceiling in choosing the closest possible world. If participants were simply employing non-counterfactual conditional reasoning, we would expect more non-CPWC responses, since similarity to the actual world would not play a role in their responses.

These results indicate that participants tend to follow the CPWC (and therefore are aligned with predictions from Hiddleston (2005)) in scenarios where there is only one closest possible world. All FORWARD scenarios only generate one closest possibility, and the single-cause scenario also only generates one closest possibility. The forward CF findings are somewhat unexpected in light of Experiment 4 in Rips (2010), where participants gave similar responses to forward and backward counterfactuals in a deterministic causal system. The backtracking results for our single-cause scenario are also novel given prior experimental work, indicating that adults are parsimonious in the changes they allow from the actual world in some backtracking cases.

We now return to our original question: *When adults reason backwards over possibilities, do they tend to take a parsimonious approach to changes in the causal system?* We find evidence that in simple cases involving just one closest possibility and its alternative (i.e., a possibility which requires more changes from the actual world), adults are in fact parsimonious over the changes they allow. They choose the closest possibility to a degree significantly above chance, mirroring our results for forward counterfactuals. However, when the selection of worlds is expanded, and adults must reason over a larger set of closest possible worlds, they no longer prefer a world from that set. Instead, they are at chance in selecting a world involving more changes from reality. This finding is rather puzzling, given the strong preference for the closest possible world in the other conditions tested.

A possible explanation could refer to the added processing difficulty involved in reasoning backwards about an effect with multiple causes. Participants must hold both causes in mind when reasoning backwards over the multi-cause scenarios, and either cause can result in the antecedent effect occurring (as opposed to the FORWARD conditions, where only one cause is mentioned in the antecedent, or in the single-cause condition where only one variable is causal). Since both causes are salient, it is possible that participants are willing to change both, thus resulting in a greater degree of non-CPWC responses.

Another possible explanation is that participants are generating the causal network differently in the case of backtracking counterfactuals. Both causes result in the antecedent event in the multi-cause scenarios, and due to the similarity between causes it is possible that some participants are treating them as a single causal unit (i.e., concluding that either both blocks are on, or both blocks are off). Further experimental manipulations to the present conditions (e.g., toggling the language and causal structures of our scenarios) would be needed to determine the validity of either explanation.

Conclusion

Backtracking remains a largely understudied aspect of counterfactual reasoning. Although prior experimental work (e.g., Rips (2010), Dehghani et al. (2012), Han et al. (2014)) has made significant strides in understanding this major component of counterfactual thought, it is still unclear from prior work the nature of the changes participants allow to their models of the world when reasoning backwards. In the present study, we take a step towards addressing this open issue, through the use of causal models within the context of Hiddleston (2005)'s theory of similarity of worlds.

We show that participants prefer models with minimal changes from the present state of affairs in some cases: namely, in forward-tracking scenarios, and in backtracking scenarios where only one cause leads to an effect. However, participants deviate from this strategy in scenarios involving multiple causes resulting in an effect. Our results provide preliminary evidence that, while adults do make use of a notion of similarity when backtracking, they are not consistent in applying this reasoning strategy. Our findings expand upon prior results in Rips (2010) with novel data, and help shed light on the degree to which participants allow changes from the actual state of affairs when reasoning counterfactually.

Acknowledgements

We would like to thank the members of the Language and Learning Lab at the University of Toronto for their helpful comments on our study materials and their ongoing discussion on the topic. This research was supported by an Insight grant from the Social Sciences and Humanities Research Council of Canada (SSHRC) to Patricia A. Ganea.

References

Danks, D. (2014). Unifying the mind: Cognitive representations as graphical models.

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

Frisch, M. (2005). Counterfactuals and the past hypothesis. *Philosophy of Science*, *72*(5), 739–750.

German, T. P., & Nichols, S. (2003). Children's counterfactual inferences about long and short causal chains. *Developmental Science*, *6*(5), 514–523.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Cogsci.*

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, *111*(1), 3.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, *71*(5), 1205–1222.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, *37*(5), 620.

Han, J.-H., Jimenez-Leal, W., & Sloman, S. (2014). Conditions for backtracking with counterfactual conditionals. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

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

Khoo, J. (2016). Backtracking counterfactuals revisited. *Mind*, *126*(503), 841–910.

Lassiter, D. (2018). Causation and probability in indicative and counterfactual conditionals.

Lewis, D. (1973). Counterfactuals and comparative possibility. In *Ifs: Conditionals, belief, decision, chance and time* (pp. 57–85). Springer.

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

Lucas, C., & Kemp, C. (2012). A unified theory of counterfactual reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Lucas, C., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological review*, *122*(4), 700.

Mackie, J. L. (1980). *The cement of the universe: A study of causation*. Clarendon Press.

Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, *132*(3), 419.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition*, *183*, 57–66.

Pearl, J. (2009). *Causality*. Cambridge University Press.

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

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–1135.

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

Spellman, B. A., & Mandel, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science*, *8*(4), 120–123.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in social psychology*, *1*, 49–72.

Ward, K. S. (2014). *Backtracking and have to: Maintaining a unified analysis of conditionals*. Unpublished doctoral dissertation, UCLA.