Cognition 209 (2021) 104574

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility

Felipe De Brigard [a,b,c,d,*], Paul Henne [e,f], Matthew L. Stanley [b,c]

[a] Department of Philosophy, Duke University, Durham, NC 27708, United States of America
[b] Department of Psychology and Neuroscience, Duke University, Durham, NC 27708, United States of America
[c] Center for Cognitive Neuroscience, Duke University, Durham, NC 27708, United States of America
[d] Duke Institute for Brain Sciences, Duke University, Durham, NC 27708, United States of America
[e] Department of Philosophy, Lake Forest College, Lake Forest, IL 60045, United States of America
[f] Neuroscience Program, Lake Forest College, Lake Forest, IL 60045, United States of America

ARTICLE INFO

Keywords:
Imagination
Counterfactual thinking
Plausibility
Similarity
Possible worlds

ABSTRACT

People frequently entertain counterfactual thoughts, or mental simulations about alternative ways the world could have been. But the perceived plausibility of those counterfactual thoughts varies widely. The current article interfaces research in the philosophy and semantics of counterfactual statements with the psychology of mental simulations, and it explores the role of perceived similarity in judgments of counterfactual plausibility. We report results from seven studies ($N = 6405$) jointly supporting three interconnected claims. First, the perceived plausibility of a counterfactual event is predicted by the perceived similarity between the possible world in which the imagined situation is thought to occur and the actual world. Second, when people attend to differences between imagined possible worlds and the actual world, they think of the imagined possible worlds as less similar to the actual world and tend to judge counterfactuals in such worlds as less plausible. Lastly, when people attend to what is identical between imagined possible worlds and the actual world, they think of the imagined possible worlds as more similar to the actual world and tend to judge counterfactuals in such worlds as more plausible. We discuss these results in light of philosophical, semantic, and psychological theories of counterfactual thinking.

1. Introduction

People frequently imagine alternative ways in which the world could have been—that is, they engage in *counterfactual thinking* (Byrne, 2016; De Brigard & Parikh, 2019; Roese & Epstude, 2017). A core feature of these kinds of thoughts is that people take some of them to be more plausible than others (Byrne, 2002; Hofstadter, 1979; Phillips, Luguri, & Knobe, 2015). For example, it seems more plausible that MySpace could have been the largest social media platform in 2020 than that the internet could have been invented in the middle ages, and this seems more plausible than the counterfactual in which humans evolved the capacity to fly at the speed of light. Judgments of counterfactual plausibility are not restricted only to fantasy, however. Many critical decisions that affect thousands of people's lives are influenced by the perceived plausibility of certain counterfactuals. For instance, several regulations issued by the Occupational Safety and Health Act and the National Safety Council in the US cover "near misses" as conditions in which accidents could have plausibly occurred but did not (NCS.org). Similarly, in tort law, cases of negligence and foreseeability are frequently settled on the basis of what is judged to have possibly happened, and often litigation involves disagreements as to whether or not it is reasonable to assume that a particular counterfactual event could have plausibly occurred (Harper, 1932).

Despite its importance, the psychological mechanisms that underlie and influence the perceived plausibility of counterfactuals remain unclear. Prior work shows that certain characteristics of the alternatives—e.g., representativeness (Kahneman & Tversky, 1972), outcome closeness (Kahneman & Tversky, 1982), event type (Byrne & McEleney, 2000), controllability (Girotto, Legrenzi, & Rizzo, 1991), and moral permissibility (Phillips & Cushman, 2017)—affect which counterfactuals people perceive as being more or less plausible. However, these various findings tend to be constrained by the particular

[*] Corresponding author at: Department of Philosophy, Duke University Durham, 203A West Duke Building, Durham, NC 27708, United States of America.
E-mail address: felipe.debrigard@duke.edu (F. De Brigard).

https://doi.org/10.1016/j.cognition.2020.104574
Received 19 April 2020; Received in revised form 19 November 2020; Accepted 22 December 2020
Available online 11 January 2021

counterfactual options included in the experimental design, so they offer only a limited understanding of the circumstances under which participants judge a counterfactual alternative as more plausible than another (Byrne, 2005; Kahneman & Miller, 1986; Petrocelli, Percy, Sherman, & Tormala, 2011). In the current study, we move beyond these approaches and investigate the role of an underexplored psychological mechanism—internal attentional allocation during the mental simulation of counterfactual thoughts—on subsequent judgments of counterfactual plausibility. In particular, our studies bring together two hitherto disparate areas of research—philosophical theory about the semantics of counterfactuals (Lewis, 1973) and recent empirical findings in the psychology of mental simulation (Schacter, Benoit, De Brigard, & Szpunar, 2015)—in order to explore whether attending to features of imagined possible situations that are either the same as or different from the actual world influences judgments of counterfactual plausibility.

Possible world semantics is a philosophical theory developed almost contemporaneously by Stalnaker (1968) and Lewis (1973). This theoretical framework suggests that the semantic evaluation of a particular conditional statement such as "If dinosaurs were alive, there would be fewer humans" is in part a matter of, first, considering possible worlds in which both the antecedent (i.e., "Dinosaurs are alive") and the consequent ("There are fewer humans") occur and, second, assessing how similar such possible worlds are to the actual world. Critically, these possible worlds are thought to be ordered by a similarity metric: some possible worlds are more similar to the actual world than others (Lewis, 1973). Thus, counterfactual statements are seen as more or less plausible depending on how similar the possible world in which they are thought to occur is to the actual world.

While the value of this metric for philosophy is undeniable (Menzel, 2019), many have criticized the notion of similarity for being vague or for leading to counterintuitive interpretations of certain types of subjunctive conditionals (Bennett, 1974; Fine, 1975; Tooley, 2002). Specifically, some have argued that there may not be a single metric along which to organize all imagined possible worlds, as our judgments of similarity may be context-dependent (Ippolito, 2016). However, others—including Lewis (1973) himself—thought that there actually may be consistency in the way people judge similarities across possible worlds, as well as in the ways in which contextual effects systematically influence such judgments (Lewis, 1999). Consistent with this idea, recent developments in philosophy and linguistics have proposed that when people consider counterfactual statements, the context in which the particular counterfactual is imbedded may highlight different aspects of the imagined possible world that is considered when evaluating the truth of the counterfactual (Ippolito, 2016), which in turn may influence the degree to which people consider a certain possible world as being more or less similar to the actual world (Greco, 2016; Lewis, 2016).

Inspired by these philosophical and linguistic ideas, the current study empirically explores the relationship between participants' *perceived* similarity across imagined possible worlds and judgments of counterfactual plausibility. Although possible world semantics was introduced as a theory for a particular purpose within philosophical logic, many thought it lacked "psychological reality" (Partee, 1977); to some, it seemed impossible to imagine a concrete possible world, let alone an infinite number of them, whenever people evaluate a counterfactual statement. But other theorists offered less cognitively taxing interpretations of possible world semantics. Stalnaker (1986), for instance, thought of possible worlds as abstractions, useful heuristics people employ to understand the meanings of statements that do not refer to reality. Others argued that even this characterization was too broad and that when assessing the possibility of a counterfactual statement, people only need to imagine a very limited subset of possible objects and relations (Barwise & Perry, 1983; Zalta, 1993). This interpretation of possible world semantics is not only more psychologically tractable, but it also dovetails seamlessly with recent developments in the cognitive science of mental simulation. Specifically, we take these imagined possible worlds to be mental simulations in which alternatives to actual facts are represented and whose contents are generated from semantic and episodic information (Irish, Addis, Hodges, & Piguet, 2012; Schacter et al., 2015). Such mental simulations, being constructed and held in working memory, are susceptible to shifts in internal attention such that different aspects of their representational content can be highlighted and prioritized (Chun, Golomb, & Turk-Browne, 2011; Myers, Stokes, & Nobre, 2017).

Against this backdrop, we attempt to determine, first, whether how plausible people think a counterfactual situation is relates to how similar they think the possible world in which the imagined situation occurs is to the actual world. Our second objective is to further investigate the cognitive processes underlying people's judgments of perceived similarity across possible worlds and its relation to judgments of counterfactual plausibility. Inspired by the aforementioned developments in the linguistics and philosophy of counterfactual statements, according to which contextual differences may highlight different aspects of the imagined possible worlds, we sought to investigate whether manipulating participants' attention to similar, versus different, aspects of the imagined possible world affects their judgments of counterfactual plausibility. If counterfactual thinking involves juxtaposing knowledge of an actual situation (i.e., in the actual world) and a mental simulation of an alternative one (i.e., in a possible world) in which the counterfactual event occurs (Byrne, 2016; De Brigard, Szpunar, & Schacter, 2013), then attending to features of the counterfactual simulation that are either the same or different from the actual one should influence the degree to which the imagined possible world is perceived as more or less similar to the actual world. Specifically, we hypothesized that attending to ways in which an imagined possible world in which a counterfactual event obtains would *differ* from the actual world would shift people's attention toward dissimilarities between the imagined possible world and the actual world, leading them to judge the counterfactual as *less* plausible. Conversely, attending to ways in which an imagined possible world in which a counterfactual event obtains would *be the same* as the actual world may shift people's attention toward similarities between the imagined possible world and the actual world, leading them to judge the counterfactual as *more* plausible. We address these general questions in the following seven studies.

2. Experiment 1a

In Experiment 1a, we sought to assess the relationship between perceived similarity of imagined possible worlds and judgments of counterfactual plausibility. To that end, we asked participants to read counterfactual statements and to imagine possible worlds in which such statements could be true. Then, for each counterfactual statement, we asked them to assess how similar the possible world in which said counterfactual could occur is to the real world, and how plausible they think the counterfactual is. As per our hypothesis, we expected a strong, positive relation between judgments of perceived similarity and counterfactual plausibility.

2.1. Methods

2.1.1. Participants

All participants in the seven experiments included in this study were recruited via Amazon's Mechanical Turk, and recruitment was restricted to individuals in the United States with a prior approval rating above 85%. All participants in all studies reported being fluent English speakers and received monetary compensation at a rate of $9/h. All the experiments reported in this paper were self-paced. A total of 100 participants were recruited in Experiment 1a, but one participant was excluded for failing to answer all the questions in the task. The final sample size and demographic information for Experiment 1a were: $N = 99$, $M_{\text{age}} = 32.34$, $SD = 8.90$, $\text{range}_{\text{age}} = [20–69]$, 46 females, 53 males. All experiments in the current study were approved by the Duke

University Campus Institutional Review Board.

2.1.2. Materials and procedure

We selected 24 statements depicting concrete and imaginable counterfactual events that violated either physical (12 statements; e.g., “Fire that freezes when touched”) or biological (12 statements; e.g., “A kitten that hatches from an egg”) laws (see Supplemental Material). Participants were then presented with all 24 counterfactual statements, one at a time, and were asked to imagine a possible world in which the counterfactual statement was true. That is, they were asked to imagine a possible world in which the counterfactual situation referred to by the counterfactual statement was actually the case. After imagining the counterfactual, participants were asked to rate how similar this imagined possible world is to the real world in a scale from 1 (*exactly the same as the real world*) to 9 (*very dissimilar from the real world*) and how plausible it is that the counterfactual statement could have been true in the real world on a scale from 1 (*very plausible*) to 9 (*very implausible*). The order in which the counterfactuals were presented was randomized across participants.

2.1.3. Statistical analyses

In all studies reported herein, data were analyzed using R (R Development Core Team, 2009) with the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2014). Data were fit to linear mixed-effects models (LMEMs). Significance for fixed effects was assessed using Satterthwaite approximations to degrees of freedom, and 95% confidence intervals around beta-values were computed using bootstrapping (*n* simulations = 1000) to estimate effect sizes (Luke, 2017). The alpha level for all statistical tests was set at 0.05.

2.2. Results

For Experiment 1a, ratings of similarity and plausibility were highly correlated across counterfactuals (Fig. 1A). Incorporating all 24 counterfactuals, a linear mixed-effects regression (LMER) of similarity on plausibility was computed, and participant and counterfactual were included as crossed random effects (with random slopes the models failed to converge, so we only modeled random intercepts). This analysis revealed a significant effect of similarity on plausibility ($b = 0.70$, $SE = 0.02$, $t = 39.20$, $p < .0001$, 95% CI = [0.66, 0.73]) such that counterfactual statements thought to be true in imagined possible worlds perceived to be more similar to the real world were also judged to be more plausible. The same pattern of results was obtained when analyzing the relationship between similarity and plausibility for biological ($b = 0.64$, $SE = 0.03$, $t = 24.97$, $p < .0001$, 95% CI = [0.59, 0.70]) and physical ($b = 0.71$, $SE = 0.02$, $t = 28.89$, $p < .0001$, 95% CI = [0.66, 0.75]) counterfactuals in separate models.

3. Experiment 1b

The results of Experiment 1a suggest a strong relation between ratings of similarity and possibility across diverse counterfactual statements. To corroborate these results and to ensure that the effects were not unique to the set of counterfactual statements employed, we conducted Experiment 1b, with an identical procedure as Experiment 1a, but using different counterfactual statements (see Supplemental Material).

3.1. Methods

3.1.1. Participants

One hundred participants were recruited and no participants were excluded. The sample size and demographic information for Experiment 1b were: $N = 100$, $M_{\text{age}} = 35.22$, $SD = 11.52$, range$_{\text{age}}$ = [20–70], 41 females, 59 males. The same recruitment specifications as in Experiment 1a applied for Experiment 1b.

3.1.2. Materials and procedure

The procedure was identical to Experiment 1a, except that a different set of counterfactual statements was employed (see Supplemental Material).

3.2. Results

Replicating the results of Experiment 1a, ratings of similarity and plausibility were also strongly associated across counterfactuals (Fig. 1B). A LMER of similarity on plausibility was computed, and participant and counterfactual were included as crossed random effects

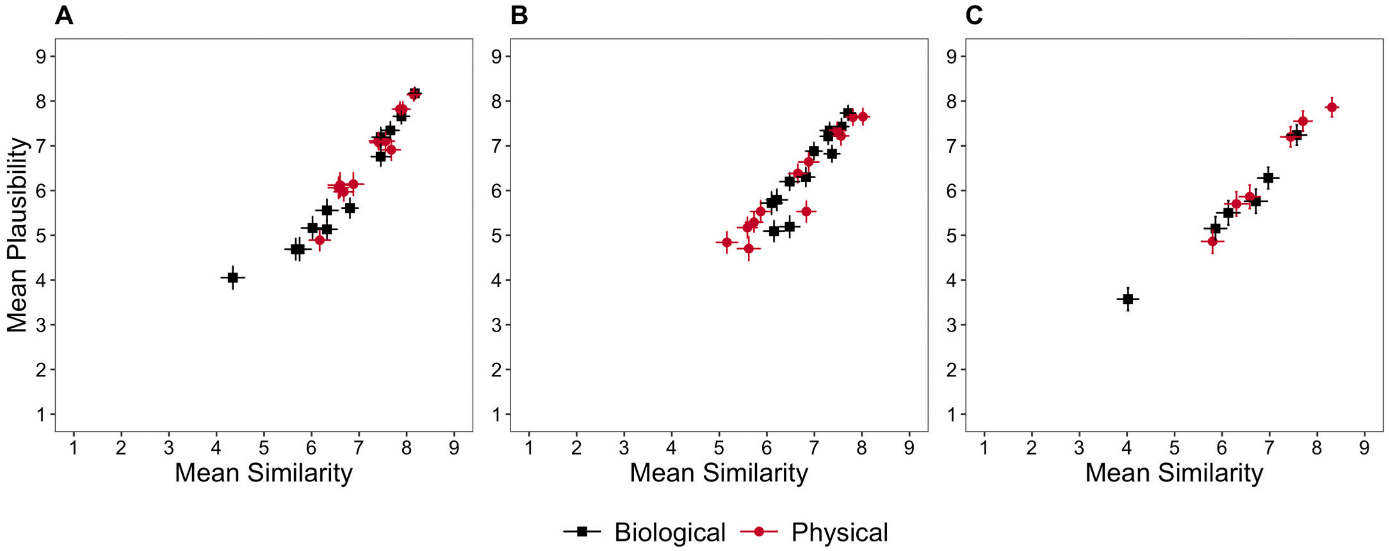

Fig. 1. Scatterplots for Experiment 1a, 1b, and 1c. Each point indicates a particular counterfactual. Black square points correspond to biological counterfactuals, and red circle points correspond to physical counterfactuals. In all three plots, the error bars represent the SEM for each counterfactual across participants. These correlation plots are for display purposes only, as they only average across counterfactual cases and do not take into account subject variability. Individual regression plots for each counterfactual statement that do incorporate individual subject variability can be found in Supplemental Materials. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(with only random intercepts included in the model). This analysis revealed a significant effect of similarity on plausibility ($b = 0.71$, $SE = 0.02$, $t = 37.38$, $p < .0001$, 95% CI = [0.67, 0.75]) such that counterfactuals thought to be true in imagined similar possible worlds were also judged to be more plausible. The same pattern of results was obtained when analyzing the relationship between similarity and plausibility for biological ($b = 0.68$, $SE = 0.03$, $t = 25.09$, $p < .0001$, 95% CI = [0.63, 0.73]) and physical ($b = 0.71$, $SE = 0.03$, $t = 26.56$, $p < .0001$, 95% CI = [0.66, 0.76]) counterfactuals in separate models.

4. Experiment 1c

Since in both experiments 1a and 1b participants judged similarity prior to plausibility, we sought to minimize possible carry-over effects that could influence the relationship between these two judgments. Thus, in Experiment 1c, similarity judgments were made in one block, and plausibility judgments were made in a separate block. The order in which these two blocks were presented was randomized across participants. The two blocks were separated by a brief unrelated distractor task. We also included filler counterfactual statements that were present in either the similarity judgments block or the plausibility judgments block, but not both. Filler items were included to help conceal the aim of the experiment. As per our hypothesis, we expected a strong, positive relationship between judgments of perceived similarity and counterfactual plausibility.

4.1. Methods

4.1.1. Participants

One hundred participants were recruited, and we excluded no participants. Sample size and demographic information for Experiment 1c were: $N = 100$, $M_{\text{age}} = 38.78$, $SD = 10.96$, $\text{range}_{\text{age}} = [22–70]$, 38 females, 61 males. The same recruitment specifications from Experiment 1a and 1b applied, except we restricted recruitment to individuals in the United States with at least 1000 completed HITs and a prior approval rating above 97%.

4.1.2. Materials and procedure

A total of 12 target statements from the lists employed in Experiments 1a and 1b were randomly selected. Of these, 6 referred to counterfactual events that violated physical laws, and the other 6 referred to counterfactual events that violated biological laws. These 12 target statements were randomly presented in both the similarity judgments block and the plausibility judgments block. Additionally, we selected 24 filler statements; 12 of these filler statements (6 biological, 6 physical) were presented only in the similarity judgments block, and the other 12 filler statements (6 biological, 6 physical) were presented only in the plausibility judgments block. All filler statements were also selected from the lists used in Experiments 1a and 1b (see Supplemental Material).

The order in which the blocks were presented was randomized across participants. In both blocks, participants were presented with 24 counterfactual statements (12 target, 12 filler), one at a time, and they were asked to imagine a possible world in which the counterfactual statement was true. That is, as in Experiments 1a and 1b, they were asked to imagine a possible world in which the counterfactual situation referred to by the counterfactual statement was actually the case. In the similarity judgments block, participants were asked to rate how similar this imagined possible world is to the real world on a scale from 1 (*exactly the same as the real world*) to 9 (*very dissimilar from the real world*). In the plausibility judgments block, participants were asked to rate how plausible it is that the counterfactual statement could have been true in the real world on a scale from 1 (*very plausible*) to 9 (*very implausible*). All participants were presented with the same brief unrelated distractor task between the similarity judgments block and the plausibility judgments block. Upon completion, participants were monetarily compensated for their time.

4.2. Results

Ratings of similarity and plausibility were highly related across counterfactuals (Fig. 1C). Incorporating all 12 target counterfactuals, a LMER of similarity on plausibility was computed, and participant and counterfactual were included as crossed random effects (with only random intercepts modeled). This analysis revealed a significant effect of similarity on plausibility ($b = 0.59$, $SE = 0.03$, $t = 21.65$, $p < .0001$, 95% CI = [0.54, 0.65]) such that counterfactual statements thought to be true in imagined possible worlds perceived to be more similar to the real world were also judged to be more plausible. The same pattern of results was obtained when analyzing the relationship between similarity and plausibility for biological ($b = 0.58$, $SE = 0.04$, $t = 16.38$, $p < .0001$, 95% CI = [0.50, 0.65]) and physical ($b = 0.57$, $SE = 0.04$, $t = 13.74$, $p < .0001$, 95% CI = [0.49, 0.66]) counterfactuals in separate models.

Taken together, the results of Experiments 1a, 1b, and 1c provide evidence of a strong relationship between judgments of counterfactual plausibility and the extent to which the imagined possible world in which the counterfactual statement is thought to be true is similar to the actual world. As such, these results speak to the first aim of the current study, and they corroborate the hypothesis that how plausible people think a counterfactual situation is relates to how similar they think the possible world in which the imagined situation occurs is to the actual world.

5. Experiment 2a

Next, we sought to experimentally manipulate perceived similarity by drawing participants' attention to features of the imagined possible world that would be either the same or different to the actual world. We reasoned that by highlighting dissimilarities—as opposed to similarities—between the actual and the imagined possible worlds, participants would judge the imagined possible world as more dissimilar to the actual world. We hypothesized that if perceived similarity influences judgments of counterfactual plausibility, when participants are shown a counterfactual statement and are asked to focus on features that would be *dissimilar* between the possible world in which the counterfactual is true and the actual world in which the counterfactual is false, they will judge the counterfactual situation as less plausible than if they attend instead to features that would be the *same* between the two worlds.

5.1. Methods

5.1.1. Participants

One-hundred twenty participants were recruited. Two participants were excluded for failing to answer all the questions in the task. As such, the sample size and demographic information for Experiment 2a were: $N = 118$, $M_{\text{age}} = 35.01$, $SD = 11.32$, $\text{range}_{\text{age}} = [19–70]$, 52 females, 65 males. The same recruitment specifications as in the previous experiments applied for Experiment 2a.

5.1.2. Materials and procedure

We presented participants with a total of 12 counterfactual statements (6 biological, 6 physical) randomly selected from the list employed in Experiment 1a, one at a time, and asked them to imagine a possible world in which the counterfactual statement was true. Next, in a within-subject design, participants generated either four ways in which the imagined world would be the same as the actual world if the counterfactual statement were true (i.e., *same* condition), or four ways that the world would be different if the counterfactual statement were true (i.e., *different* condition). Six of the 12 counterfactuals were assigned to the *same* condition, while the other 6 were assigned to the *different* condition. Counterfactual assignment was fully counterbalanced across participants. Finally, as in Experiments 1a and 1b, participants were asked to give a rating of similarity (1 - *exactly the same as the real world* to 9 - *very dissimilar from the real world*) and a rating of plausibility (1 - *very plausible* to 9 - *very implausible*).

5.2. Results

To investigate whether there were differences in similarity ratings as a function of condition (same versus different), a LMER of condition on similarity was computed, and participant and counterfactual were included as crossed random effects (random intercepts only). This revealed a significant effect of condition ($b = 1.33$, $SE = 0.10$, $t = 13.45$, $p < .0001$, 95% CI = [1.14, 1.53]). When participants generated four ways in which the imagined possible world in which the counterfactual statement is true would be the *same* as the real world, they tended to think of them as being more similar to the real world relative to when they generated four ways in which the imagined possible world would be *different* from the real world (Fig. 2). The same pattern of results was obtained when analyzing the effect of condition on similarity for biological ($b = 1.9$, $SE = 0.13$, $t = 10.18$, $p < .0001$, 95% CI = [1.04, 1.54]) and physical ($b = 1.36$, $SE = 0.14$, $t = 9.54$, $p < .0001$, 95% CI = [1.09, 1.66]) counterfactuals in separate models. These results suggest that the manipulation was successful in affecting participants' judgments of perceived similarity of imagined possible worlds.

Next, to investigate whether there were differences in plausibility ratings as a function of condition (same versus different), a LMER of condition on plausibility was computed, and participant and counterfactual were included as crossed random effects (random intercepts only). This revealed a significant effect of condition on plausibility ($b = 0.69$, $SE = 0.11$, $t = 6.37$, $p < .0001$, 95% CI = [0.47, 0.90]). When participants generated four ways in which the imagined possible world in which the counterfactual statement is true would be the *same* as the real world, they tended to think of the counterfactual as being more plausible than when they generated four ways in which the imagined possible world would be *different* from the real world. The same pattern of results was obtained when analyzing the effect of condition on similarity for biological ($b = 0.68$, $SE = 0.15$, $t = 4.61$, $p < .0001$, 95% CI = [0.38, 0.98]) and physical ($b = 0.69$, $SE = 0.15$, $t = 4.67$, $p < .0001$, 95% CI = [0.40, 0.99]) counterfactuals in separate models.

Finally, to investigate whether the perceived similarity of the counterfactuals in each condition mediated the effect on plausibility, we conducted mixed-effects mediation analyses to examine the average causal mediation effect (ACME) and the average direct effect (ADE). The results revealed that while the ACME was significant, ACME = 0.87, $p < .001$, 95% CIs [0.70, 1.04], the ADE was not, ADE = −0.17, $p = .09$, 95% CI = [−0.37, 0.04], suggesting that participants' perception of the similarity of the imagined possible world mediated the relationship between condition (*same*, *different*) and their judgment of counterfactual plausibility (Fig. 2).

6. Experiment 2b

Employing a within-subject design, Experiment 2a manipulated perceived similarity by asking participants to imagine possible worlds where they thought a particular counterfactual statement would be true, and to focus on aspects of such imagined worlds that would either be the same or different to the actual world. Our results suggest not only that the manipulation affected participants' judgments of perceived similarity and counterfactual plausibility, but also that the shift in perceived similarity mediated the effect of condition on judgments of counterfactual plausibility. To corroborate this result, control for any possible order effects due to the within-subject design, and more thoroughly examine the role of perceived similarity in judgments of counterfactual

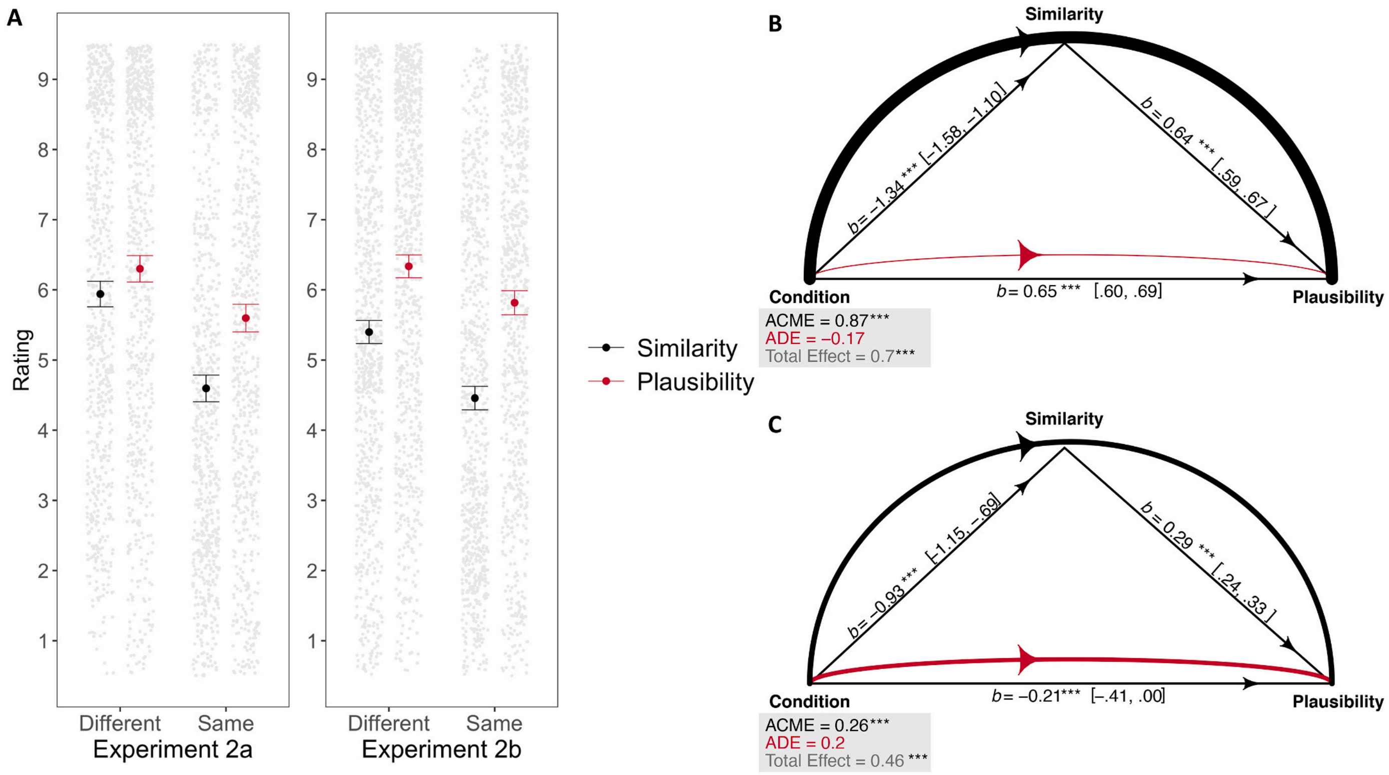

Fig. 2. (*A*) Mean rating of similarity and plausibility as a function of condition in Experiment 2a and 2b collapsed across items. Bars indicate 95% confidence intervals. Light grey points represent individual data points evenly jittered. (*B*) Mediation model described in Experiment 2a. The width of the curved arrows representing the average direct effect (ADE) and average causal mediation effect (ACME) lines are weighted by the proportion of effect mediated in the model. 95% confidence intervals in brackets. (*C*) Mediation model described in Experiment 2b. The width of the curved arrows representing the average direct effect (ADE) and average causal mediation effect (ACME) lines are weighted by the proportion of effect mediated in the model. 95% confidence intervals in brackets.

plausibility, we conducted a between-subject version of Experiment 2a.

6.1. Methods

6.1.1. Participants

The same recruitment strategy and specifications as in the previous experiments apply to Experiment 2b. Since this between-subject experiment involves 2 conditions and 12 counterfactual statements, with each individual participant being randomly assigned to only one of them, we increased the sample size so that we would have around $n = 70$ per counterfactual statement, after expected exclusions. As such, we recruited a total of 1700 participants; 62 participants failed to answer all the questions in the task, so they were removed. As such, the sample size and demographic information for Experiment 2b were: $N = 1638$, $M_{\text{age}} = 35.69$, $SD = 12.00$, $\text{age}_{\text{range}} = [18, 87]$; 738 females, 852 males.

6.1.2. Materials and procedure

Experiment 2b used the same counterfactual statements as Experiment 2a, but now each participant was randomly assigned to either a *same* or *different* condition, with each participant seeing *only one* counterfactual statement. The assignment of counterfactual statements was counterbalanced across participants.

6.2. Results

To investigate differences in similarity ratings as a function of condition, a LMER of condition on similarity was computed, revealing a significant effect of condition ($b = 0.93$, $SE = 0.12$, $t = 7.88$, $p < .0001$, 95% CI = [0.69, 1.17]). As in Experiment 2a, participants who generated four ways in which the imagined possible world in which the counterfactual statement is true would be the *same* as the real world, tended to think of them as being more similar to the real world relative to those participants that generated four ways in which the imagined possible world would be *different* from the real world (Fig. 2). The same pattern of results emerged when analyzing the effect of condition on similarity for biological ($b = 1.03$, $SE = 0.17$, $t = 6.17$, $p < .0001$, 95% CI = [0.72, 1.39]) and physical ($b = 0.82$, $SE = 0.16$, $t = 5.00$, $p < .0001$, 95% CI = [0.51, 1.17]) counterfactuals in separate models.

Next, to investigate whether there were differences in plausibility ratings as a function of condition, a LMER of condition on plausibility was computed, revealing a significant effect of condition ($b = 0.47$, $SE = 0.11$, $t = 4.19$, $p < .0001$, 95% CI = [0.25, 0.68]). Once again, as in Experiment 2a, when participants generated four ways in which the imagined possible world in which the counterfactual statement is true would be the same as the real world, they tended to think of the counterfactual as being more plausible than when they generated four ways in which the imagined possible world would be *different* from the real world. The same pattern of results emerged when analyzing the effect of condition on similarity for biological ($b = 0.40$, $SE = 0.16$, $t = 2.52$, $p = .012$, 95% CI = [0.08, 0.69]) and physical ($b = 0.54$, $SE = 0.16$, $t = 3.45$, $p = .0006$, 95% CI = [0.23, 0.86]) counterfactuals in separate models.

Finally, to investigate whether the perceived similarity of the counterfactuals in each condition mediated the effect on plausibility, we conducted mixed-effects mediation analyses. This analysis revealed that the ACME was significant, ACME = 0.26, $p < .001$, 95% CIs [0.18, 0.34], and the ADE was not, ADE = 0.20, $p = .07$, 95% CI = [0.00, 0.41], suggesting that participants' perception of the similarity of the imagined possible world mediated the relationship between condition (*same*, *different*) and their judgment of counterfactual plausibility. Taken together, the results from Experiment 2b corroborated, in a between-subjects design, the within-subject results from Experiment 2a, and lend further support to the hypothesis that judgments of counterfactual plausibility are influenced by how similar people think the possible worlds in which counterfactual events occur are to the actual world.

7. Experiment 3a

The results from Experiments 2a and 2b indicated that manipulating participants' perception of the similarity between the imagined possible world in which a counterfactual statement is thought to be true, and the actual world in which the counterfactual statement is known to be false, affects their judgments of how plausible it is that such counterfactual events could have occurred. To gain further evidence in favor of this claim, we conducted Experiment 3a, which followed the same logic as Experiment 2a. In this experiment, we included an additional condition in which participants generated two ways in which the imagined possible world would be different from the actual world *and* two ways in which it would be the same (i.e., *both* condition). If the previous manipulation is truly affecting participants' perception of perceived similarity, then asking participants to focus on *both* similarities and differences between the actual world in which the counterfactual event does not occur and the imagined possible world in which the event occurs should lead their ratings to fall between those in the *same* and the *different* conditions for both perceived similarity and counterfactual plausibility.

7.1. Methods

7.1.1. Participants

The same recruitment strategy and specifications as in the previous experiments apply to Experiments 3a. Since this experiment involves 3 conditions and 18 counterfactual statements per participants, we increased the sample size so that we would have around $n = 100$ per condition after expected exclusions. As such, a total of 310 participants were recruited; four participants were excluded as they failed to complete the task. As such, the sample size and demographic information were: $N = 306$, $M_{\text{age}} = 33.95$, $SD = 8.91$, $\text{age}_{\text{range}} = [19, 70]$; 138 females, 161 males.

7.1.2. Materials and procedures

Experiment 3a followed the same within-subject structure as Experiment 2a, except for two differences: 1) it included a randomly selected subset of 18 of the 24 counterfactuals used in Study 1a, and 2) there were now three conditions: *same*, *different* and *both*. Participants received six counterfactuals in each of the three conditions. This assignment was counterbalanced across participants.

7.2. Results

We computed LMER models of condition (*same*, *different*, and *both*) on similarity, and participant and counterfactual were included as crossed random effects (random intercepts only). This analysis revealed that when participants generated four ways in which the world would be the same, the worlds in which the counterfactuals obtain are judged to be more similar to the real world compared with counterfactuals where participants generated two ways in which the world would be the same and two ways in which the world would be different ($b = 0.53$, $SE = 0.06$, $t = 8.83$, $p < .0001$, 95% CI = [0.41, 0.66]). By contrast, when participants generated four ways in which the world would be different, the worlds in which the counterfactuals obtain are judged to be less similar to the real world compared with counterfactuals where participants generated two ways in which the world would be the same and two ways in which the world would be different ($b = -0.47$, $SE = 0.06$, $t = 7.79$, $p < .0001$, 95% CI = [−0.59, −0.35]). Finally, and replicating the results from Experiment 2a, when participants generated four ways in which the imagined possible world would be the same as the actual, the worlds in which the counterfactuals obtain are judged to be more similar to the real world compared with counterfactuals where participants generated four ways in which the world would be different ($b = 1.01$, $SE = 0.06$, $t = 16.62$, $p < .0001$, 95% CI = [0.88, 1.14]; Fig. 3).

LMER models of condition on plausibility (as before, participant and

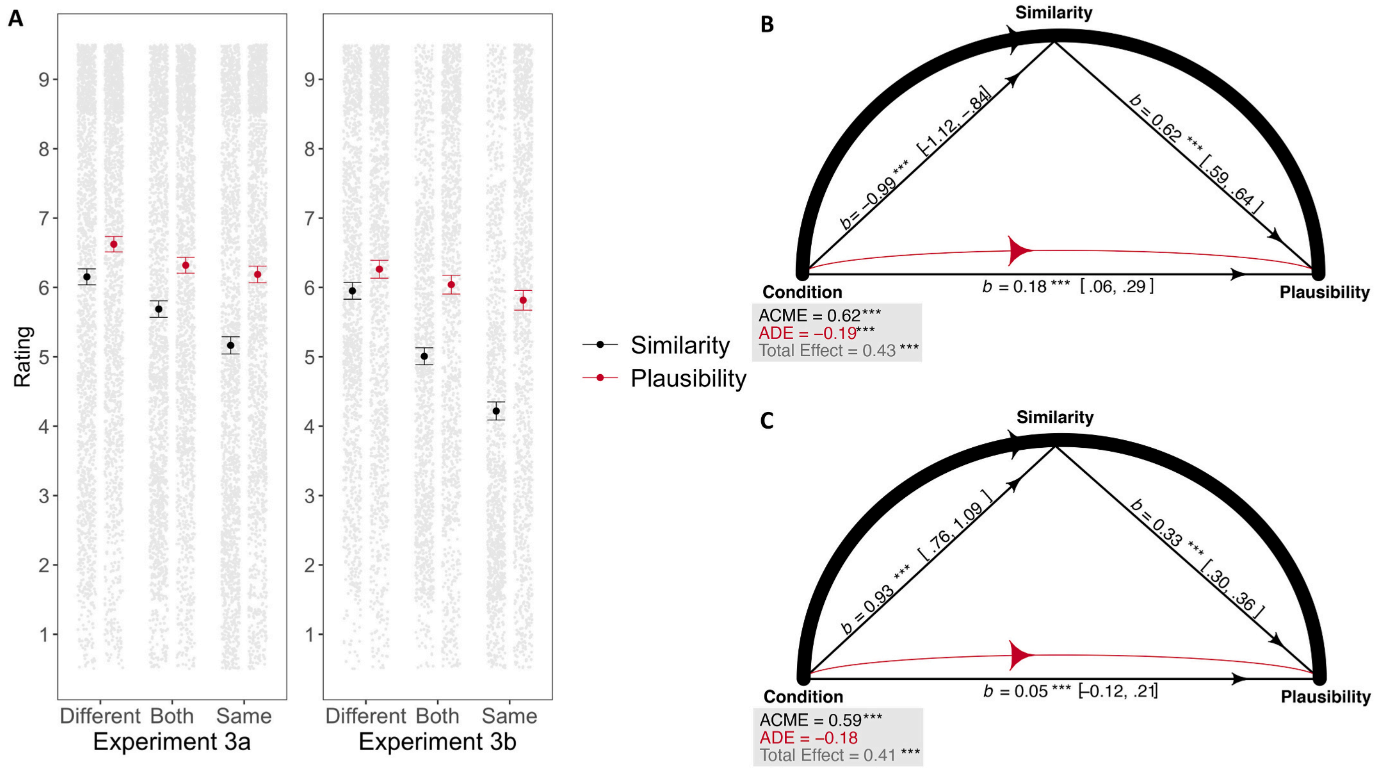

Fig. 3. (*A*) Mean rating of similarity and plausibility as a function of condition in Experiment 3a and 3b collapsed across items. Bars indicate 95% confidence intervals. Light grey points represent individual data points evenly jittered. (*B*) Mediation model described in Experiment 3a. The width of the curved arrows representing the average direct effect (ADE) and average causal mediation effect (ACME) lines are weighted by the proportion of effect mediated in the model. 95% confidence intervals in brackets. (*C*) Mediation model described in Experiment 3b. The width of the curved arrows representing the average direct effect (ADE) and average causal mediation effect (ACME) lines are weighted by the proportion of effect mediated in the model. 95% confidence intervals in brackets.

counterfactual were modeled as crossed random effects with random intercepts only) also revealed that when participants generated four ways in which the imagined possible world would be the same, the relevant counterfactual statement was judged to be more plausible relative to when participants generated two ways in which the imagined possible world would be the same and two ways in which it would be different ($b = 0.15$, $SE = 0.06$, $t = 2.39$, $p = .017$, 95% CI = [0.03, 0.27]). By contrast, when participants generated four ways in which the imagined possible world would be different, the relevant counterfactual statement was judged to be less plausible relative to when participants generated two ways in which the imagined possible world would be the same and two ways in which it would be different ($b = -0.31$, $SE = 0.06$, $t = -4.93$, $p < .0001$, 95% CI = [−0.43, −0.19]). Lastly, and replicating the results from Experiment 2a, when participants generated four ways in which the imagined possible world would be the same as the actual, the relevant counterfactuals are thought to be more plausible compared to when participants generated four ways in which the imagined possible world would be different ($b = 0.46$, $SE = 0.06$, $t = 7.31$, $p < .0001$, 95% CI = [0.34, 0.58]; Fig. 3).

Finally, to investigate whether the perceived similarity of the counterfactuals in each condition mediated the effect on plausibility, we conducted mixed-effects mediation analyses. This analysis revealed that both ACME and ADE were significant when comparing the same and different conditions, ACME = 0.62, $p < .001$, CI = [0.53, 0.71]; ADE = −0.19, $p < .001$, 95% CI = [−0.28, −0.09] (Fig. 3). Once again, these results indicate that participants' perception of the similarity of the imagined possible world partially mediated the relationship between condition (*same*, *different,* and *both*) and their judgment of counterfactual plausibility.

8. Experiment 3b

To corroborate the results from Experiment 3a, control for any possible order effects due to the within-subject design, and to further examine the role of perceived similarity in judgments of counterfactual plausibility, we conducted a between-subject version of Experiment 3a.

8.1. Methods

8.1.1. Participants

The same recruitment strategy and specifications as in the previous experiments apply to Experiment 3b. Since this between-subject experiment involves 3 conditions and 18 counterfactual statements, with each individual participant being randomly assigned to only one of them, we increased the sample size so that we would have roughly $n = 75$ per counterfactual statement, after exclusions. As such, we recruited a total of 4100 participants. Fifty-six participants were excluded, as they failed to answer all the questions in the task. As such, the sample size and demographic information were: $N = 4044$, $M_{\text{age}} = 34.74$, $SD = 11.36$, $\text{age}_{\text{range}} = [18, 99]$; 2043 females, 1964 males.

8.1.2. Materials and procedure

Experiment 3b followed the same logic as Experiment 3a, but (as Experiment 2b) it employed a between-subject design, such that each participant was randomly assigned to either the *same, different,* or *both* condition, and each saw only one counterfactual statement. This assignment was counterbalanced across participants.

8.2. Results

LMER models of condition (with counterfactual modeled as a random effect with random intercepts only) on similarity showed that when participants generated four ways in which the imagined possible world would be the same as the actual world, the possible world in which the counterfactual obtains is judged to be more similar to the real world compared with counterfactuals where participants generated two ways in which the possible world would be the same and two ways in which the possible world would be different ($b = 0.78$, $SE = 0.09$, $t = 9.04$, $p < .0001$, 95% CI = [0.60, 0.96]). By contrast, when participants generated four ways in which the imagined possible world would be different from the actual world, the possible world in which the counterfactual obtains is judged to be less similar to the real world compared with counterfactuals where participants generated two ways in which the possible world would be the same and two ways in which the possible world would be different ($b = -0.93$, $SE = 0.09$, $t = 10.94$, $p < .0001$, 95% CI = [−1.10, −0.76]). Finally, and replicating the results from Experiment 2b, when participants generated four ways in which the imagined world would be the same as the actual world, the possible world in which the counterfactual obtains is judged to be more similar to the real world compared with counterfactuals where participants generated four ways in which the possible world would be different ($b = 1.71$, $SE = 0.09$, $t = 19.98$, $p < .0001$, 95% CI = [1.54, 1.87]; Fig. 3).

Next, LMER models of condition on plausibility revealed that when participants generated four ways in which the imagined possible world would be the same, the relevant counterfactual statement was judged to be more plausible relative to when participants generated two ways in which the imagined possible world would be the same and two ways in which it would be different ($b = 0.22$, $SE = 0.09$, $t = 2.40$, $p = .017$, 95% CI = [0.06, 0.40]). By contrast, when participants generated four ways in which the imagined possible world would be different, the relevant counterfactual statement was judged to be less plausible relative to when participants generated two ways in which the imagined possible world would be the same and two ways in which it would be different ($b = -0.19$, $SE = 0.09$, $t = 2.11$, $p = .035$, 95% CI = [−0.36, −0.01]). Lastly, and replicating the results from Experiment 2b, when participants generated four ways in which the imagined possible world would be the same as the actual, the relevant counterfactuals are thought to be more plausible compared to when participants generated four ways in which the imagined possible world would be different ($b = 0.41$, $SE = 0.09$, $t = 4.51$, $p < .0001$, 95% CI = [0.24, 0.58]; Fig. 3).

Finally, to investigate whether the perceived similarity of the counterfactuals in each condition mediated the effect on plausibility, we conducted mixed-effects mediation analyses. This analysis revealed that the ACME was significant when comparing the same and different conditions, but the ADE was not, ACME = 0.59, $p < .001$, CI = [0.50, 0.67]; ADE = −0.18, $p = .06$, CI = [−0.35, 0.00]. The significant ACME indicates that participants' perceived similarity mediated the relationship between condition and their judgment of counterfactual plausibility.

9. General discussion

When we imagine alternative ways the world could have been, we tend to think of some of those alternatives as more plausible than others. But which psychological factors make people think that certain counterfactual thoughts are more plausible than others remains unclear. The current study explored a concrete suggestion inspired by philosophical research in the semantics of counterfactual statements as well as cognitive research on mental simulation. The starting point, in brief, is that when people evaluate the plausibility of a counterfactual statement, they entertain a mental simulation of a possible world in which the counterfactual situation occurs, and then they assess the relative similarity between such a possible world and the actual world (Lewis, 1973). Accordingly, the more dissimilar to the actual world people perceive the imagined possible world to be, the less plausible they would think the counterfactual is. The results from Experiments 1a 1b, and 1c suggest that this may be the case, at least for a set of counterfactual statements violating biological and physical laws. Using both between- and within-subject designs, we found a strong correlation between judgments of counterfactual plausibility and the perceived similarity of the possible world in which the relevant counterfactual event occurs and the actual world.

The results of these first three experiments share a strong family resemblance with prior research on the representative heuristic, or the tendency to judge the likelihood of an event based upon how much it resembles a stereotypical representative of the relevant population (Kahneman & Tversky, 1972). For instance, in a seminal study, Kahneman & Tversky (1973) asked participants to read about a fictional character, Tom W., who was described as having certain features thought to be prototypical of a graduate student. One group of participants judged the likelihood that Tom W. was a graduate student (as opposed to several other possible professions), whereas a different group of participants were asked to judge how similar was Tom W. to the prototypical graduate student. Both ratings were very tightly correlated ($r = 0.97$). Similar correlations were found for other characters and professions (Kahneman & Frederick, 2001; Tversky & Kahneman, 1983). Consistent with these findings, the results of Experiments 1a, 1b, and 1c suggest that perceived similarity and plausibility are strongly correlated during counterfactual thinking. However, our results extend beyond the representative heuristic framework by showing an effect of perceived similarity, not in relation to a representative instance of the relevant target population, but rather in relation to the imagined possible world in which a counterfactual event is thought to obtain and the actual world.

To further explore the connection between perceived similarity of imagined possible worlds and judgments of counterfactual plausibility, we employed an experimental manipulation aimed at shifting participants' attention during the generation of the mental simulations that gave content to their counterfactual thoughts. The inspiration behind this task came from a combination of two recent lines of research. First, developments in modal semantics suggest that the linguistic context in which counterfactual statements are embedded influences not only our assessments of their truth-value but also the perceived plausibility of the counterfactual situations to which they refer (Greco, 2016; Ippolito, 2016; Lewis, 2016). Second, consistent with the idea that counterfactual thoughts involve the generation and maintenance of mental simulations in working memory (Byrne, 2016; Johnson-Laird, 1983; Schacter et al., 2015), we sought to capitalize on the finding that shifts in internal attentional allocation can highlight different aspects of the mentally simulated content (Chun et al., 2011; Myers et al., 2017). Accordingly, we reasoned that by asking participants to focus their attention on differences between the actual and the imagined possible world, they may be more likely to deem such a possible world as being more dissimilar to the actual one, relative to when they are asked to focus their attention on similarities between the actual and the imagined possible worlds. Using both between- and within-subject designs and multiple stimulus sets, the results of Experiments 2a, 2b, 3a, and 3b lend support to this hypothesis. Moreover, the results of these experiments also support the claim that these shifts in perceived similarity between the actual and the imagined possible worlds systematically influence our judgments of counterfactual plausibility. Specifically, the results of Experiments 2a and 2b lend credence to the hypothesis that when participants perceive a possible world in which a counterfactual is thought to occur as more dissimilar to the actual world, they judge said counterfactual as less plausible than when they think of the possible world as more similar to the actual world, in which case they think of the relevant counterfactual as more plausible. The results of Experiments 3a and 3b not only replicate those of Experiments 2a and 2b (both between- and within-subjects), but they also suggest that the relationship between perceived similarity and counterfactual plausibility may be linear, such that when participants attended to both similarities *and* differences between the actual and the possible world in which a counterfactual statement is thought to occur, their judgments of counterfactual plausibility fell between those made when participants attended only to similarities or only to differences.

9.1. Theoretical implications

A pressing conceptual question at this point is how to understand and interpret the notion of an imagined possible world. Albeit frequently used in philosophy and modal semantics, both philosophers and semanticists disagree about how to interpret this concept. Lewis (1973), for instance, thought of possible worlds as concrete real entities that could confer truth values to counterfactual statements. By contrast, Stalnaker (1986) thought of them as abstractions, useful heuristics people employ to understand the meaning of statements that do not refer to a concrete reality. For some researchers, however, both accounts are psychologically doubtful, for it is unlikely that one imagines a *whole* possible world when assessing the possibility of a counterfactual statement, rather than a very limited subset of possible objects and relations or *situations.* (Barwise & Perry, 1983; Zalta, 1993). We believe that this *situationist* interpretation of the notion of an imagined possible world in philosophy and semantics is consistent with the psychological notion of *mental simulation*, understood as a dynamic mental representation entertained in working memory that iconically represents possible items and their relationships as constrained by prior knowledge (Craik, 1943).

The idea that mental simulations support counterfactual thinking was initially elaborated within the *mental models* framework (Johnson-Laird, 1983; Byrne, 2002, 2005; but see also Hegarty, 2004) and more recently has received an alternative computational interpretation within the context of causal reasoning in the *counterfactual simulation model* (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2020; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017). Our suggestion is that this notion of mental simulation dovetails with recent developments in the psychology and neuroscience of memory and imagination, according to which mental simulations are thought of as episodic-like experiences, the building blocks of which come from stored information in episodic memory, constrained by knowledge accumulated in semantic memory (Schacter et al., 2015). Accordingly, to entertain a possible world in a situationist sense is to mentally generate a dynamic simulation in working memory, whose episodic content iconically represents the arrangement of possible objects and the relations among them. However, these simulations need not be fully developed, in the sense that they can be coarse and represent items sketchily with poor resolution and few details. Internal attention, however, can highlight different aspects of the mental simulation, and thus prioritize them for further processing (Myers et al., 2017). Building upon this idea, we argue that drawing attention to similarities or differences between episodic simulations of possible worlds (i.e., situations) in which an event could have occurred, relative to episodic simulations constrained by our knowledge of the actual world, prioritizes different information in a way that influences people's subsequent judgments of counterfactual plausibility.

Although underexplored, some researchers have investigated the role of attention in counterfactual thinking. It has been found, for instance, that drawing attention to information about the background leading up to an event when thinking about what could have occurred instead, makes people more likely to consider certain counterfactual alternatives (Kahneman & Tversky, 1982; Seelau, Seelau, Wells, & Windschitl, 1995). Our results are consistent with these findings. But why does shifting attention from similarities to differences in imagined possible worlds decreases the extent to which people perceive the counterfactuals as plausible? A possible explanation comes from the recent computational model of plausibility judgments known as the *Plausibility Analysis Model* (PAM; Connell and Keane, 2006). According to this model, when assessing the plausibility of a situation, people evaluate the conceptual coherence of the features of the imagined possibility and their previous knowledge. When the features of the imagined items are incoherent with those in the actual world, people judge the imagined simulation as less plausible than when the features of the imagined items cohere with those in the actual world. Our results suggest that, at least for the case of counterfactual thinking, it may be useful to think of the notion of coherence in terms of perceived similarity. More precisely, we suggest that it may be useful to extend the notion of conceptual coherence to iconic theories of similarity, such as matching-based accounts (e.g., Tversky, 1977), whereby a mental representation of a certain situation A is judged to be more similar to B than C if the simulated features of A match better those of B than of C. As such, when our attention is focused on features of the imagined possible words (i.e., situations) that match those of our knowledge of the actual world better, we tend to think of such worlds as more similar to ours—and the counterfactual statements that obtain in them as more plausible—than when our attention is focused on features that do not match, or match less well, with what we know of the actual world. Extending the notion of conceptual coherence in the PAM to cover matching-based accounts of similarity is, we think, a useful avenue for future research.

9.2. Limitations, future research and conclusions

The literature on counterfactual thinking also suggests a number of limitations in our studies. We know, for instance, that background knowledge affects whether or not people consider certain possibilities as more or less likely (Wells, Taylor, & Turtle, 1987). As such, we suspect that our results would also vary as a function of expertise. Maple-syrup producing bees, for instance, may be deemed more implausible to expert entomologists than to amateurs, who in turn may think of such possible worlds as more similar to ours than expert entomologists do. We would expect expertise to moderate our effects on an item-level. Additionally, it may be possible that attending to essential versus accidental similarities between the known items in the actual world and the imagined ones in the possible world could interact with our judgments of plausibility. If this is the case, it may be possible that our results could differ when the physical or biological violation affects essential versus accidental features. And we also know that statistical (Kahneman & Miller, 1986) and moral norms (Phillips & Cushman, 2017) constrain our counterfactual selection (Bear & Knobe, 2017). As such, perceived similarity and plausibility may differ when the laws violated by counterfactual situations are social or moral than when they are physical or biological. Clearly, further research would be needed to help disentangle these intricate questions.

The results of experiments 2 and 3 also indicate that the experimental manipulation had a greater effect on judgments of similarity than it did on judgments of counterfactual plausibility. This difference is notable for at least two reasons. First, it corroborates that, while related, participants' ratings of perceived similarity and counterfactual plausibility do not tap onto the same underlying factor. Second, it also suggest that while the perceived similarity between the actual and the imagined possible world influence our judgments of counterfactual plausibility, there are other factors that must account for the difference in the effect size. While our results do not tell us what these factors may be, future research may be able to shed light on this issue. We also believe that the hypothesis offered here opens up avenues for future research as well. To account for the results reported in our experiments, we have marshalled an explanatory hypothesis based upon recent developments in philosophy (Greco, 2016; Lewis, 2016), linguistics (Ippolito, 2016), and the cognitive psychology (e.g., Gerstenberg et al., 2017; Gerstenberg et al., 2020) and neuroscience (e.g., De Brigard & Parikh, 2019; Irish et al., 2012; Schacter et al., 2015) of mental simulations. Nevertheless, it is possible that alternative hypotheses that do not involve the postulation of dynamic mental simulations in working memory or shifts on internal attention could offer a better fit for the data. We hope that future research could help to further test the strength of our hypothesis.

Another limitation of our studies is that they are confined to individual counterfactual statements, and do not directly examine more complex expressions, such as subjunctive counterfactual conditionals of the form "If A had been the case, then B would have been the case" (Von Fintel, 2012). Nevertheless, we believe that our findings could illuminate how people assess the antecedent—and, possibly, the consequent—of a counterfactual conditional. After all, if our Lewis-inspired proposal is on the right track, when considering the meaning of a statement serving as, say, the antecedent in a counterfactual conditional, individuals may need to imagine first a possible world against which to evaluate their truth value. If such processes are susceptible to attentional manipulations of the sorts reported here, it may be worth exploring whether or not they also systematically affect judgments of plausibility for the counterfactual conditional in which such individual statements are embedded.

As we mentioned above, researchers have identified a number of difficulties with Lewis' proposal (e.g., Blumson, 2018; Fine, 1975; Kroedel & Huber, 2013; Morreau, 2010). Indeed, Lewis himself was aware of some (Goodman, 1970), and he anticipated others (Lewis, 1973: 91). Nevertheless, he thought that even if there were difficulties with the formalization of comparative similarity, as a matter of psychological fact people may think of comparative similarity across possible worlds in a systematic—or, in his words, "coordinated"—way (Lewis, 1973: 92). This suspicion inspired our studies, which is why we sought to measure participants' *perceived* similarity between an actual and an imagined possible world—that is, how similar they *think* these two worlds are—rather than *actual* similarity, i.e., how similar the actual and the imagined possible world are according to some objective measure. As a result, we refrain from making straightforward conclusions from perceived to actual similarity in the modal domain. Moreover, given that differences in perceived versus actual similarity have also been documented in other domains, including social and personality psychology (Montoya, Horton, & Kirchner, 2008) as well as perceptual cognition (Lowet, Firestone, & Scholl, 2018), it remains an open question how our results on the perceived similarity across imagined possible worlds bear on formal models of comparative similarity in counterfactual contexts. Since some of these formal models are geared toward counterfactual conditionals, such as Pearl's structural counterfactual approach (Pearl, 2000, 2013) or Petrocelli et al.'s (2011) model of counterfactual potency, further research employing more complex counterfactual expressions would be needed. Nevertheless, the fact remains that perceived similarity between the actual and imagined possible worlds seems to play a role in participant's judgments of counterfactual plausibility in single statements.

There are certainly many factors that influence the precise contours of the situations we choose to consider when mentally simulating alternatives (Phillips, Morris, & Cushman, 2019), and within those, many different factors constrain our preferences to think of one or another counterfactual event (Seelau et al., 1995). The results of the studies reported herein contribute to the literature on counterfactual constraints by providing evidence in favor of the claim that perceived similarity plays an important role in shaping people's judgment of plausibility during counterfactual thinking. Moreover, our results suggest that shifts in internal attention that result in changes of perceived similarity influence people's judgment of counterfactual plausibility. Accordingly, our findings not only lend credence to recent philosophical and linguistic views according to which contextual differences influence our understanding of counterfactual statement, but they also suggest that manipulating the interpreter's attention toward features that are more or less similar to the actual world influences the degree to which they think of counterfactuals as more or less plausible. Exploring the repercussions of this attentionally-dependent contextual shift in counterfactual judgments not only for philosophy but legal context is, we think, a fruitful avenue for future research. Undoubtedly, a better understanding of how people make judgments of counterfactual plausibility, and the factors that may systematically influence such judgments, has the potential to inform occupational and legal outcomes affecting millions of people in the real world.

CRediT authorship contribution statement

Felipe De Brigard: Conceptualization, Formal Analysis, Funding Acquisition, Methodology. **Paul Henne:** Conceptualization, Data curation, Formal Analysis, Methodology. **Matthew L. Stanley:** Conceptualization, Data curation, Formal Analysis, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Paul Bello, Sangeet Khemlani, Tobias Gerstenberg, Louie T. D. Henne, Laura Niemi, Carlotta Pavese, Jonathan Phillips, Walter Sinnott-Armstrong, the members of the Imagination and Modal Cognition lab at Duke University, and the audience of the Summer School *Thinking about the Possible* at the Central European University, in Budapest, for their feedback. This research was supported by a grant from the Office of Naval Research (N00014-17-1-2603) to FDB.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104574.

References

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge: MIT Press.
Bates, D. M, Mäechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.
Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167*, 25–37.
Bennett, J. (1974). Counterfactuals and possible worlds. *Canadian Journal of Philosophy, 4*, 381–402.
Blumson, B. (2018). Distance and dissimilarity. *Philosophical Papers, 48*(2), 211–239.
Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Science, 6*(10), 426–431.
Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge: MIT Press.
Byrne, R. M. J. (2016). Counterfactual thought. *Annual Review of Psychology, 67*, 135–157.
Byrne, R. M. J., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1318–1331.
Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62*, 73–101.
Connell, L., & Keane, M. T. (2006). A model of plausibility. *Cognitive Science, 30*, 95–120. https://doi.org/10.1207/s15516709cog0000_53.
Craik, K. J. W. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science, 28*(1), 59–66.
De Brigard, F., Szpunar, K. K., & Schacter, D. L. (2013). Coming to grips with the past: Effect of repeated simulation on the perceived plausibility of episodic counterfactual thoughts. *Psychological Science, 24*(7), 1329–1334.
Fine, K. (1975). Review of David Lewis's "counterfactuals". *Mind, 84*, 451.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2020). A counterfactual simulation model of causal judgment. in review https://psyarxiv.com/7zj94/.
Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731–1744.
Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica, 78*, 111–133.
Goodman, N. (1970). Seven strictures on similarity. In L. Foster, & J. W. Swanson (Eds.), *Experience and theory*. MA: University of Massachusetts Press.
Greco, D. (2016). Safety, explanation, iteration. *Philosophical Issues, 26*, 187–208.
Harper, F. V. (1932). Foreseeability factor in the law of torts. *Notre Dame Law Review, 7* (4), 468–482.
Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences, 8*(6), 280–285.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. NY: Vintage.

Ippolito, M. (2016). How similar is similar enough? *Semantics and Pragmatics, 9*(6), 1–60.

Irish, M., Addis, D. R., Hodges, J., & Piguet, O. (2012). Exploring the content and quality of episodic future simulations in semantic dementia. *Neuropsychologia, 50*, 3488–3495.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Kahneman, D., & Frederick, S. (2001). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136–153.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kahneman, D., & Tversky, A. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–210). New York: Cambridge University Press.

Kroedel, T., & Huber, F. (2013). Counterfactual dependence and arrow. *Noûs, 47*(3), 453–466.

Lewis, D. (1973). *Counterfactuals*. Oxford: Oxford University Press.

Lewis, D. (1999). Elusive knowledge. In D. Lewis (Ed.), *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

Lewis, K. S. (2016). Elusive counterfactuals. *Nous, 50*(2), 286–313.

Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics, 80*, 1278–1289.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*, 1494–1502.

Menzel, C. (2019). Modal set theory. In O. Bueno, & S. Shalkowski (Eds.), *The Routledge handbook of modality*. NY: Routledge.

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships, 25*(6), 889–922.

Morreau, M. (2010). It simply does not add up: Trouble with overall similarity. *Journal of Philosophy, 107*(9), 469–490.

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in Cognitive Sciences, 21*, 449–461.

Partee, B. H. (1977). Possible world semantics and linguistic theory. *The Monist, 60*(3), 303–326.

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

Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive Science, 37*, 977–985.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology, 100*(1), 30–46.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences, 114*(18), 469–4654.

Phillips, J., Luguri, J., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition, 145*, 30–42.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences, 23*(12), 1026–1040.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new controversies, new insights. *Advances in Experimental Social Psychology, 56*, 1–79.

Schacter, D. L., Benoit, R., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory, 117*, 14–21.

Seelau, E. P., Seelau, S. M., Wells, G. L., & Windschitl, P. D. (1995). Counterfactual constraints. In N. J. Roese, & J. M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp. 57–79).

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*. Oxford: Oxford University Press.

Stalnaker, R. (1986). Possible worlds and situations. *Journal of Philosophical Logic, 15*, 109–123.

Tooley, M. (2002). Backward causation and the Stalnaker-Lewis approach to counterfactuals. *Analysis, 62*(3), 191–197.

Tversky, A. (1977). Features of similarity. *Psychological review, 84*(4), 327–352.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

Von Fintel, K. (2012). Subjunctive conditionals. In G. Russell, & D. Graff Fara (Eds.), *The Routledge companion to philosophy of language* (pp. 466–477). NY: Routledge.

Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology, 53*, 421–430.

Zalta, E. N. (1993). Twenty-five basic theorems in situation and world theory. *Journal of Philosophical Logic, 22*(4), 385–428.