Manuscript version accepted for publication in the Journal of Experimental Psychology: Learning, Memory, and Cognition
DOI: https://dx.doi.org/10.1037/xlm0001354

Reasoning about Actual Causation in Reversible and Irreversible Causal Structures

Simon Stephan
Department of Psychology
University of Göttingen
Gosslerstrasse 14
37073 Göttingen
Germany

Abstract

This paper investigates people's judgments of actual causation in the context of a previously neglected property of causal structures – their reversibility, that is, whether an effect persists or returns to its original state if its causes are removed. Causal reversibility, and its potential impact on causal judgment, was recently analyzed theoretically by Ross and Woodward (2022). They hypothesized that reversibility might affect people's evaluation of causes in late-preemption scenarios. The typical finding in preemption scenarios is that events happening earlier are considered to be actual causes, while events happening later are regarded as non-causes. The hypothesis is that this robust intuition depends on causal reversibility, and that in reversible structures later events are regarded as actual causes. Across three main experiments and one supplementary study ($N = 590$), it is shown that reversibility has the predicted effect: later causes are perceived to make an actual causal contribution to the effect. It is also shown that Henne, Perez, and McCracken (2023), in a first study, did not find evidence for Ross and Woodward's hypothesis because they did not test whether people regard later causes in preemption-like sequences of reversible structures as maintainers and not as triggers of their effect. Because they used test questions that asked explicitly for triggering rather than maintaining, or were at least ambiguous, their results seemed to show that people think that later events have no causal impact. Maintaining is a relevant causal concept deserving more attention in both philosophical theories and psychological studies on causal cognition.

Keywords: actual causation, causal reasoning, late preemption, causal reversibility, maintaining

Simon Stephan https://orcid.org/0000-0002-6557-9637

Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany. E-mail: simon.stephan@psych.uni-goettingen.de. Declarations of interest: none

All experimental materials, data, and R (R Core Team, 2022) analyses scripts (R version 4.2.2) for all main and pilot studies have been made publicly available on an OSF repository here: `https://osf.io/dnbf6/`. They can also be accessed via a GitHub page at `https://simonstephan31.github.io/actual_cause_reverse/`.

1 Introduction

Consider one of the famous examples of *redundant causation* from the philosophical and psychological literature on causation (see, e.g., Danks, 2017; Hall, 2004; Halpern, 2016; Hitchcock, 2001; Lewis, 2000; Paul & Hall, 2013; Ross & Woodward, 2022), which goes something like this: Suzy and Billy are two perfectly accurate rock throwers. Whenever either of them decides to fling a rock towards a bottle, the bottle shatters into pieces. It so happens that Suzy and Billy are aiming at the same bottle, but Suzy manages to throw her rock a split second earlier than Billy. Her rock arrives at the bottle first and hits it. The bottle shatters into pieces. When Billy's stone reaches the location where the bottle used to stand, it meets nothing but thin air. Had Suzy not thrown her stone, Billy's stone would have hit the bottle and shattered it. In this and structurally similar scenarios almost everyone agrees that the first event, Suzy's throwing her rock in the given scenario, was the actual (or singular) cause of the effect, the bottle's shattering in this case, whereas the second, Billy's throwing his rock in the example scenario, was completely non-causal (in addition to philosophical papers cited above that defend this notion see also Chang, 2009; Henne, Kulesza, Perez, & Houcek, 2021; Lombrozo, 2010; Rose & Danks, 2012; Stephan, Mayrhofer, & Waldmann, 2020; Walsh & Sloman, 2011, for psychological studies showing that lay people also share this intuition).

The story above represents a scenario of redundant causation because the effect is *overdetermined*: it would still have occurred even if the event that we intuitively judge to be its *actual cause* would not have occurred because the effect would then have been brought about by the alternative cause. More precisely, this specific scenario instantiates a situation philosophers have come to call *late preemption.* It is called "late" preemption because the alternative (second) cause[1] that would generate the effect if the first did not happen actually occurs (i.e., its causal process is already "on its way"/unfolding but gets preempted because the causal influence of the first cause "arrives at the effect" first); Billy throws his rock even though Suzy already threw hers a bit earlier (for examples of "early preemption", also called "back-up" scenarios, where the second cause occurs only if the first one does not, see, e.g., Hitchcock, 2001, 2007, 2009).

Scenarios of preemption like this one have received interest in philosophy and in psychological studies of causal reasoning (see, e.g., Chang, 2009; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Henne et al., 2021; Lombrozo, 2010; Stephan et al., 2020; Stephan & Waldmann, 2017; Walsh & Sloman, 2011) because they are apparent counterexamples to (simple) *counterfactual dependency theories* (see, e.g., Lewis, 1973, 2000) of (singular/actual) causation (see Hitchcock, 2009; Paul, 2009; Waldmann, 2017, for overviews on counterfactual theories of actual causation). According to such theories, an actual (or singular) event c is considered to be the cause of a subsequent event e if it is true that e would not have occurred if c had not occurred[2]. This principle is famously violated in cases

[1] Note that the term "cause" is used here to describe the second event in a preemption situation even though this event is considered to be non-causal in the specific situation. The reason why it is still called "cause" here is that it represents an instance of a *general* or *type-level* cause that has the *potential* to cause the effect (e.g., in the absence of the first cause).

[2] Singular, or actual, causation refers to causal relations holding between specific events that actually occur at a particular time in a particular place. By contrast, *general causation* refers to re-instantiable causal relations between types of events. An example for a general causal relation would be smoking causing

of redundant causation like preemption (for refined counterfactual theories that can handle preemption see, e.g., Halpern & Hitchcock, 2015; Halpern & Pearl, 2005a; Lewis, 1973).

In a recent theoretical article, Ross and Woodward (2022) revisited such cases because they noted that they all share a specific *structural* feature: they all play out in what Ross and Woodward (2022) have called "irreversible", or "one-hit", causal structures. The characteristic feature of such irreversible structures is that the effect (e.g., a bottle's shattering) cannot be undone or reversed (at least in a practical sense[3]). As soon as the state of the effect variable has undergone a change (e.g., form absent, 0, to present, 1), it persists. In contrast to irreversible cases, the effect variable in reversible causal structures can return to its original state (e.g., from present, 1, back to absent 0). One way this could happen is by removing its (actual) cause(s) (or, in other words, by turning them back off). Ross and Woodward (2022) consider this an interesting case because they think that our intuitions about actual causation in preemption scenarios might actually depend on whether the underlying causal structure is irreversible or reversible. From a theoretical perspective, causal reversibility is interesting because it has so far been widely neglected.

The present paper addresses the empirical question of how lay people's intuition about actual causation in preemption scenarios changes depending on causal reversibility. In standard scenarios of late preemption that play out in irreversible structures, people reliably have been found to say that only the event that occurs first makes an actual causal contribution to the effect (see, e.g., Gerstenberg et al., 2021; Henne et al., 2023; Stephan et al., 2020; Walsh & Sloman, 2011), while the one occurring second (or later in the case of more than two potential causes) is regarded as non-causal. How might reasoners' causal intuition change if such a scenario were to play out on the stage of a reversible causal structure?

1.1 How reversibility might influence causal intuitions in preemption scenarios

Ross and Woodward (2022) assume that what would change in preemption scenarios playing out in reversible structures is our intuition about the second (cause) event. It would now be regarded as exerting a causal influence on the effect, too. A crucial question is why this should be the case. Ross and Woodward (2022) notice that in reversible structures the potential causes tend to be of a different nature than in irreversible structures[4]. In typical scenarios of irreversible causation, we seem to perceive the causes of an effect as *triggers* of the effect. Therefore, our focus appears to be on the moment in which the change of the state of the effect variable occurs. In such irreversible cases, any causal influence on the effect seems to disappear as soon as the effect has occurred. For example, once Suzy has thrown her rock and the rock shattered the bottle, the rock does not continue to exert any influence on it.

heart attacks.

[3] One could imagine a world in which all the shards of a broken bottle are glued together again, but such scenarios are rather theoretical.

[4] It should be noted that Ross and Woodward (2022) consider reversible scenarios in which the effect would return to its original value if the target causes were to be removed. This can be contrasted with reversible cases in which the effect would persist if the target causes were to be removed but could be reset due to the influence of external causes. For example, a damaged blood vessel might not recover if a person stopped smoking, but could be fixed through surgery.

By contrast, in reversible structures causes not only seem to trigger an effect but also to *maintain* or *sustain* it. One example of reversibility given by Ross and Woodward (2022) is a spring getting stretched by a weight. Attaching a weight to a spring not only triggers its stretching but also maintains it. Assuming the absence of external influences, the spring does not return to its initial length until the weight is removed again. It is not hard to come up with further examples, which suggests that reversible causation is frequently found in our lives: going out in the sun triggers a tan and staying in the sun maintains it. A nasty back pain gets caused by an inflamed nerve and does not cease until the nerve gets better. The economy gets ruined by corruption and will not recover until corruption is overcome.

An abstract scheme illustrating the temporal dynamics in cases of irreversible and reversible scenarios of the kind that Ross and Woodward (2022) consider is given in Fig. 1. The upper graph (a) illustrates a scenario with an irreversible causal structure. The onset of the effect (e) follows shortly after the onset of the first cause (c_1). The effect persists even when that cause disappears again. Thus, potential causes of the effect turning on and off after this initial onset of the effect are not followed by any changes in the effect. A typical sequence of late preemption is given in this first graph by the first three events that occur at t_1, t_2, and t_3. The lower graph (b) illustrates a reversible scenario. Again, the effect onset occurs shortly after the onset of the first cause, but this time the effect turns off again as soon as one of its causes disappears, resulting in temporal *effect gaps* during periods where none of the effect's causes is present. The temporal order of events having been found to lead to preemption intuitions in irreversible cases is given in this graph by the last three events that occur at t_4, t_5, and t_6. Under a reversible structure, such a sequence of events may be referred to as a "preemption-like sequence"[5].

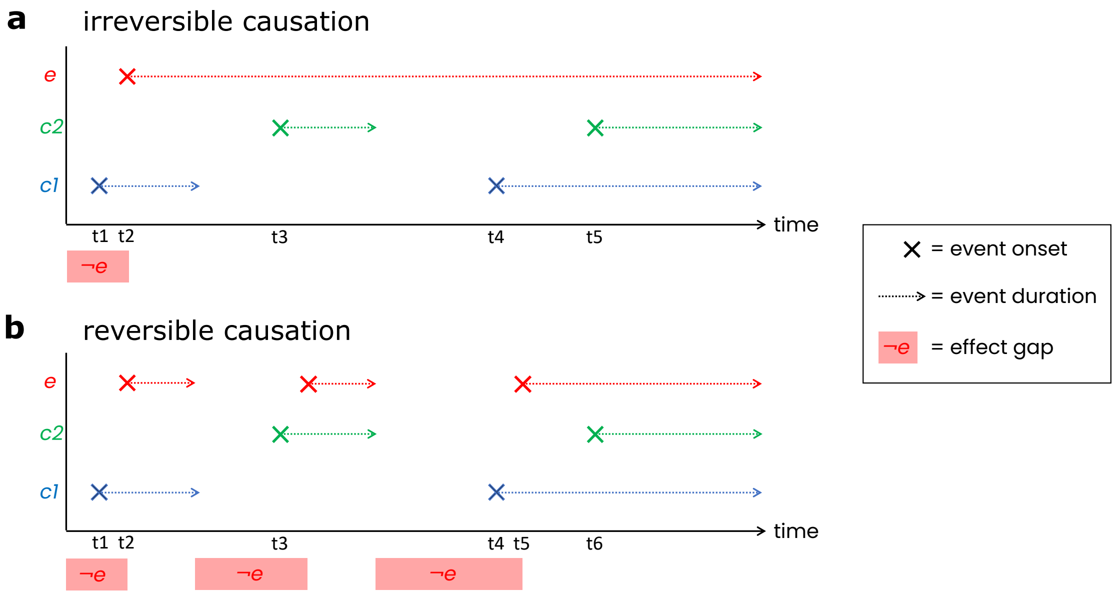
According to Ross and Woodward (2022), the maintaining aspect of reversible structures would leave room in preemption-like sequences for the second event to be perceived as making an actual causal contribution to the effect. In other words, unlike in irreversible late preemption scenarios, the second cause would no longer be regarded as non-causal. As a thought experiment contrasting preemption intuitions in reversible and irreversible cases, Ross and Woodward (2022) consider a scenario about two switches connected to a light. In the irreversible version of the scenario, turning on one of the switches would turn on the light, but turning it off again would *not* turn the light off again; once the light is on, it keeps burning. The first graph (a) in Fig. 1 may be taken to illustrate this scenario if c_1 and c_2 represent the two light switches and e represents the light. In the target situation (taking place between t_1 and t_3), one of the switches is turned on before the other. This first switch getting turned on seems to be the event causing the light to turn on, while the second switch getting turned on seems to exert no influence – it is preempted by the first. In the reversible switch scenario (which appears to be the more natural one), turning on a switch turns on the light and turning it off again turns the light back off. Here, the light keeps burning for as long as at least one switch is on. The lower graph (b) in Fig. 1 may illustrate this case. As before, in the target situation (taking place between t_4 and t_6) one switch gets turned on before the other. Ross and Woodward (2022) have the intuition that the second switch might now also be perceived as exerting a causal influence on the effect.

Importantly, although unlike in the irreversible case the second event in the re-

[5] This terminology is used because if the Ross-Woodward hypothesis turned out to be correct, there would not be the typical notion of preemption in such a sequence of a reversible causal structure.

Figure 1

Illustration of temporal dynamics in irreversible and reversible causation.



Note. *Illustration of the temporal dynamics between two causes, c1 and c_2, and an effect, e, in (a) irreversible and (b) reversible causal structures. Crosses represent event onsets (at different points in time, t_n) and arrows event durations. In the irreversible case depicted in (a), the effect occurs shortly after the first cause and then remains present. In the reversible case depicted in (b), the effect disappears again when all causes are absent. Unlike in the reversible case, this may result in "effect gaps" at later points.*

versible case might indeed be perceived as making a causal contribution, Ross and Woodward's (2022) analysis still reveals a difference between the two causes. The first cause seems to be both a *trigger* and a *maintainer* of the effect, whereas the second event only seems to be a *maintainer*. As will be shown later, this has implications for how one has to probe people's intuitions about the second cause in such sequences to reveal its perceived causal status.

1.2 Current evidence for the influence of reversibility

There is so far, to the best of the author's knowledge, only one set of empirical studies that addresses the question of how reversibility might affect people's causal intuitions in preemption scenarios. This set of studies, comprising three main studies and one supplementary study, was conducted by Henne et al. (2023). In all experiments, the authors used the light switch scenario introduced presented above. According to their own account, the authors' study that best tests the scenario is the supplementary study, in which participants read the following scenario description and were asked the following test questions:

> David designed a special light with two switches: a red switch and a blue switch. If either switch was turned on, a purple light would turn on.

> [Reversible condition:] Once the purple light goes on, it can be turned off by turning off the switches.
>
> [Irreversible condition:] Once the purple light goes on, it cannot be turned off at all – even by turning off the switches.
>
> At the exact same moment, David set timers so that the red switch and the blue switch will each turn on sometime today. At 1:00PM, the red switch turned on, so the purple light turned on. At 2:00PM, the blue switch turned on. Nothing else changed, so at 4:00PM the purple light was still on.
>
> [Test query:] To what extent do you agree with the statement about the passage you just read?
>
> [Early cause condition:] The purple light was on at 4:00PM because the red switch turned on.
>
> [Late cause condition:] The purple light was on at 4:00PM because the blue switch turned on.

Both reversibility and whether the test query was about the early or the late cause were manipulated between participants. Participants provided their actual cause judgments on a nine-point rating scale (with endpoints labeled "strongly disagree" and "strongly agree").

Contrary to Ross and Woodward's (2022) expectation, Henne et al. (2023) did not find any substantial influence of reversibility. Just like in standard irreversible late preemption scenarios, participants strongly agreed that the first event is an actual cause and they tended to disagree that the second event is an actual cause. This pattern was found in all their experiments, although there was weak descriptive trend in the expected direction in the studies. Does this mean that causal reversibility does not influence people's causal intuitions in preemption scenarios? Henne et al. (2023) were careful in the interpretation of their results, though. They report that they failed to find evidence for Ross and Woodward's hypothesis, but they did not claim to have refuted it.

1.3 The role of test query formulation and test scenario

One reason why Henne et al. (2023) might not have detected the predicted effect is that the test question about the second cause that they used in their experiments was not optimal to find an effect. In their Experiments 1 and 2, the statements for both causes that participants evaluated were "The purple light went on because David turned on the red [blue] switch". As has been described earlier, Ross and Woodward (2022) assume that the second cause in reversible structures might be regarded as a maintainer, not as a trigger. A problem might be that the test statements in Henne et al.'s (2023) Experiments 1 and 2 refer to triggering causes and not maintaining causes. Participants might thus have disagreed because the second cause in the scenario is not a trigger. This, however, does not imply that participants believed that the second cause had no causal impact on the effect.

This leads to the hypothesis that an influence of reversibility can be observed if test questions clearly ask for maintaining rather than triggering. In their Experiment 3 and in a supplementary study, the test statement about the second cause that Henne et al. (2023)

used was the one given above: "The purple light was on at 4:00PM because the blue switch turned on". As this statement is still not completely unambiguous, participants might still have expressed that the second cause did not trigger the effect.

In addition to an ambiguous test query formulation, what also might have led participants to interpret the test statement as one referring to triggering is that triggering is what seems to come to mind naturally in a scenario about light switches. If somebody gets asked what a light switch does, probably most people would intuitively respond that a light switch "turns on the light" rather than that it "keeps the light burning". In a light switch scenario the event that is responsible for the onset of the effect seems to be in the focus. This focus on the triggering capacity of light switches might result from our daily experience with them. An effect of reversibility might thus be easier to find in other scenarios.

1.4 Overview of experiments

Three experiments were conducted. Experiment 1 tests the hypothesis that in the light switch scenario an effect of reversibility on people's judgments of the second cause will be observed once the test question clearly refers to maintaining rather than triggering. To foreshadow the findings, this is what was found. The study also aims to replicate Henne et al.'s (2023) finding that no (substantial) effect of reversibility occurs if the original test statements are used.

Experiments 2a and b aim to generalize the findings beyond the light switch scenario. Experiment 2a tests a fictitious biological scenario about squids that can change their color. Experiment 2b tests a mechanical scenario that bears some resemblance to the example about the metal spring that Ross and Woodward (2022) presented to illustrate causal reversibility. Apart from testing different scenarios, a further difference of Experiments 2a and 2b from Experiment 1 is that they conveyed the relevant information with dynamic stimuli. In Experiment 2a animations were used, and in Experiment 2b subjects were shown short video clips.

1.4.1 Transparency and openness

All experimental materials, data, and R (R Core Team, 2022) analyses scripts (R version 4.2.2) for all main and pilot studies have been made publicly available in an OSF repository here: `https://osf.io/dnbf6/` (Stephan, 2024a). They can also be accessed via a GitHub page at `https://simonstephan31.github.io/actual_cause_reverse/` (Stephan, 2024e). This GitHub repository website also contains demo versions of the experiments that readers can run in their own browser. All experiments were implemented as online experiments using the *jsPsych* library (de Leeuw, Gilbert, & Luchterhandt, 2023). All studies reported in this paper were pre-registered at the OSF here `https://osf.io/dnbf6/registrations`, and the sample sizes were determined based on a priori power analyses and pilot studies. The link to each study's individual pre-registration and the sample size rationales for each experiment will be given below in the respective experiment sections.

1.4.2 Ethics

All experiments reported in this paper were conducted in accordance with the Declaration of Helsinki and the Ethical Principles of the German Psychological Society (DGPs), the Association of German Professional Psychologists (BDP), and the American Psychological Association (APA). The reported experiments involved no invasive or otherwise ethically problematic techniques and no deception; participants, prior to their participation, received information stating the goals and content of the experiments. Informed consent was obtained from all participants. For these reasons, according to national jurisdiction, a separate vote by a local institutional review board was not required (see also the regulations on freedom of research in the German Constitution (§5 (3)), and the German University Law (§22)).

2 Experiment 1

This study consisted of two parts. The first part was a conceptual replication of Henne et al.'s (2023) supplementary study, which was described above. It is a conceptual replication, because the different test statements that participants rated were manipulated within-participant in the present study, whereas Henne et al. (2023) manipulated all factors between participants. This part of the study had participants rate the acceptability of the two original test statements presented above in the theory section.

The second part was identical except for the test statements participants rated. This part tested the hypothesis that an effect of reversibility on participants' ratings for the second cause depends on test statement formulation. Participants who participated in this part of the study provided agreement ratings for four instead of two causal statements. For each cause of the scenario (the two light switches), participants rated a trigger and a maintainer statement. The crucial prediction was that participants will agree more with the claim that that the second cause is a maintainer when the causal structure is reversible than when it is irreversible. At the same time, participants in both reversibility conditions should agree that the first cause is a trigger of the effect.

The materials of this study were pre-tested in a pilot study ($N = 92$) whose results are summarized at the repository site. Experiment 1's pre-registration can be accessed here `https://osf.io/twy6c` (Stephan, 2024c). A demo version of the experiment can be run at `https://simonstephan31.github.io/actual_cause_reverse/MainExperiments/Exp1/experiment_files/task_local_demo/Exp_demo.html`.

2.1 Methods

2.1.1 Participants

One hundred and fifty-five participants took part in this experiment. The data of five participants were excluded prior to any analyses because they failed to give the correct answer to a control question (as specified in the pre-registration). The final sample thus consisted of $N = 150$ participants ($M_{age} = 40$ years, $SD_{age} = 14.22$ years, age range 18 to 75 years) recruited via the online platform `www.prolific.co` participated in this online study and provided complete data. The inclusion criteria were a minimum age of 18 years, English as first language, and an approval rate (concerning participants' participation in online studies hosted via prolific) of 90 percent. To ensure that all participants were able to understand the written instructions, prolific workers with "no formal qualifications" for the criterion "highest education level completed" were excluded from participation. Participants also were asked to take part via PC or Laptop and not via Tablet or Smartphone. Participants who took part in the pilot or other studies of this paper were not allowed to participate.

Fifty participants of this sample took part in the replication part of the experiment while the other 100 participants took part in the novel part of the study.

2.1.2 Sample size rationale

The sample sizes for both parts of the study were based on a priori power analyses. The effect sizes used in the planning were informed by what was observed in the pilot study.

For the replication part, the power analysis was based on the predicted main effect of *causal statement* tested in a mixed ANOVA conducted with R's *afex* package (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2022). The goal was to detect an effect size of $\eta_p^2 = .50$ with at least 90% test power (for further details, see the pre-registration). The power analysis showed that this level of test power would be obtained with $N = 14$. However, it was decided to obtain more data points than those that would result from only 14 participants. It was decided to have a total sample size of $N = 50$ in this part of the study (n 25 in each between-participants condition).

As for the novel part of the experiment, the basis of the sample size planning was the predicted interaction effect between *causal statement* and *causal reversibility* tested in a mixed ANOVA with R's *afex* package. As before, the goal was to reach at least 90% test power (for further details, see the pre-registration), this time for an effect size of $\eta_p^2 = .15$. The result of the analysis was a sample of $N = 82$. It was decided to realize a slightly larger sample of $N = 100$ ($n = 50$ in each between-participants condition).

2.1.3 Design, materials, and procedure

The replication part of the study had a 2 (causal structure: *irreversible* vs. *reversible*; between-participants) × 2 (causal statement: *original statement for first cause* vs. *original statement for second cause*; within-participant in random order) mixed design.

The novel part of the study had a 2 (causal structure: *irreversible* vs. *reversible*; between-participants) × 4 (causal statement: *first cause triggered* vs. *second cause triggered* vs. *first cause maintained* vs. *second cause maintained*; within-participant in random order) mixed design.

In both parts of the experiment, participants were alternately assigned to the different between-participant conditions. Before participants proceeded to the main part of the study, they had to confirm that they taking part via PC or laptop and that they were willing to pay attention.

The two parts (the replication and the novel part) of the experiment only differed with respect to the causal statements about the test scenario that participants saw and evaluated. All participants read Henne et al.'s (2023) light switch scenario. Participants in the replication part were shown both original causal statements: "The purple light was on at 4:00PM because the red switch turned on", and "The purple light was on at 4:00PM because the blue switch turned on". As in the original experiment, participants were asked to say how much they agreed with these statements. They provided their agreement ratings on nine-point rating scales (with endpoints labeled "strongly disagree" and "strongly agree").

Participants in the novel part evaluated four causal statements, which differentiated between triggering and maintaining causation. The two trigger statements were:

- The red switch turning on at 1:00PM caused the purple light to turn on.
- The blue switch turning on at 1:00PM caused the purple light to turn on.

and the two maintainer statements were:

- At 4:00PM, the red switch being on is keeping the purple light burning.
- At 4:00PM, the blue switch being on is keeping the purple light burning.

All four statements were presented in random order on the same screen as the scenario description.

After participants had provided their agreement ratings, they proceeded to a novel screen where they were asked a comprehension check question probing their understanding of the causal reversibility of the presented scenario: "In the scenario about the switches and the light you just read, once the purple light goes on, can it be turned off by turning off the switches?" Participants answered this question by selecting "Yes" or "No". Only the data of participants who responded correctly to this question were kept for the analyses.

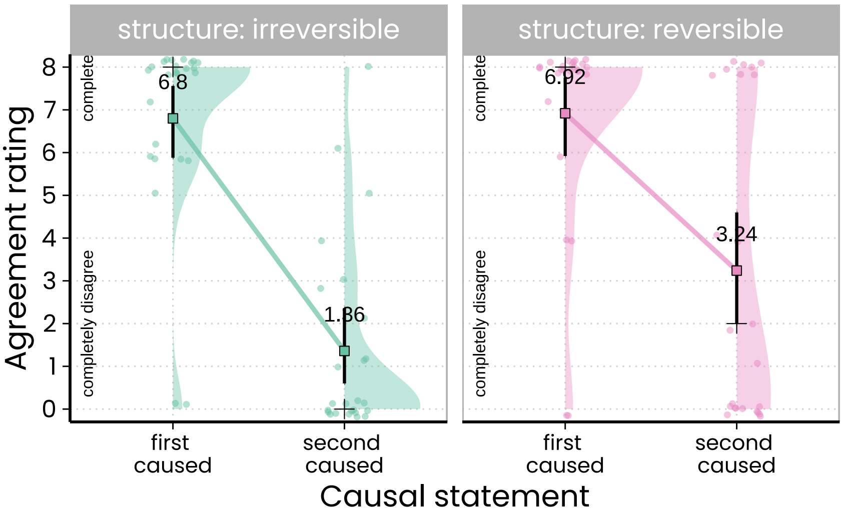
Participants then provided demographic data, were given the opportunity to report any technical errors they might have encountered, and then finished the experiment on a short debriefing screen.

2.2 Results and discussion

The results of the replication part of the study are summarized in Fig. 2, which shows participants' agreement ratings to the two different causal statements. As can be seen there, participants strongly agreed with the causal statement that the purple light was on at 04:00PM because the red (first) switch turned on. They agreed much less with the statement that the light was on at 04:00PM because the blue (second) switch turned on. It can also be seen in Fig. 2 that this was true in both the irreversible ($M_{first} = 6.80$, 95% CI $[5.87, 7.73]$; $M_{second} = 1.36$, 95% CI $[0.16, 2.56]$) as well as in the reversible scenario ($M_{first} = 6.92$, 95% CI $[5.99, 7.85]$; $M_{second} = 3.24$, 95% CI $[2.04, 4.44]$), although ratings for the second cause were slightly higher in the reversible case. This trend in the expected direction had already been observed in Henne et al.'s (2023) experiments. However, the mean rating for the second cause stayed below the midpoint of the scale, indicating that participants still tended not to regard the second cause as an actual cause of the effect. As has been noted earlier, this might have been the case because participants tended to interpret the causal statement as one that is asking for a trigger rather than for maintainer of the effect. At the same time, the half-violin plot visualizing the distribution of the ratings and the jittered dots representing participants' individual ratings also show that there was a subgroup of participants in the reversible structure condition who actually highly agreed with the causal statement about the second cause.

Figure 2

Participants' agreement ratings in the replication part of Experiment 1.



Note. Squares and annotations denote means, "+" denote medians. Error bars represent 95% confidence intervals. Jittered dots show participants' individual ratings, and the density plots their distribution.

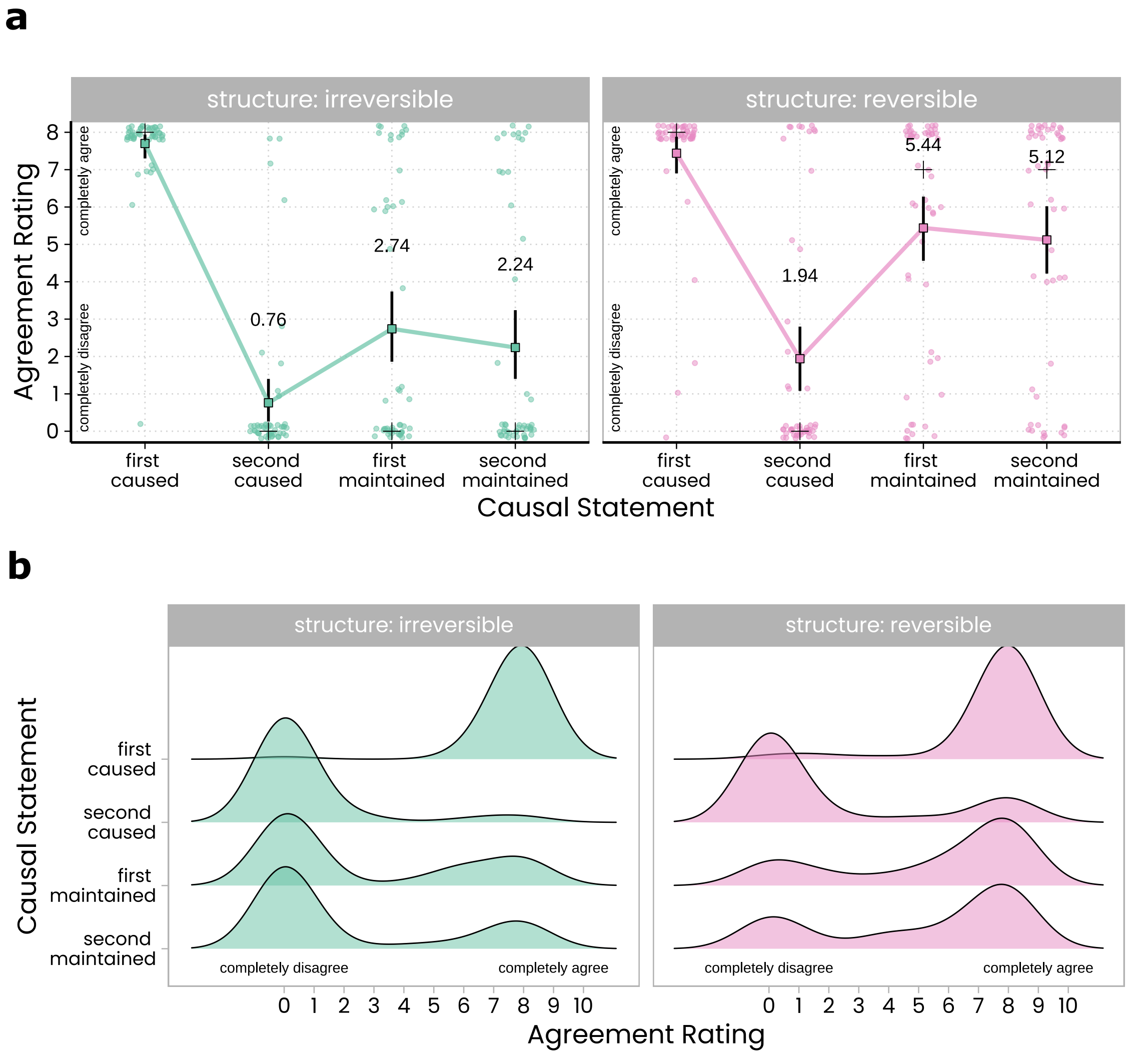
A mixed ANOVA conducted with R's *afex* package (Singmann et al., 2022) yielded a main effect of causal reversibility, $F(1, 48) = 5.90$, $p = .019$, $\eta_p^2 = .109$. This effect was obtained because ratings were on average higher in the reversible structure condition. It also yielded a significant main effect of causal statement, $F(1, 48) = 51.90$, $p < .001$, $\eta_p^2 = .520$. This main effect was obtained because ratings for the first cause were higher than for the second cause. Planned directed contrasts conducted with R's *emmeans* package (Lenth, Singmann, Love, Buerkner, & Herve, 2018) showed that this was the case in both the irreversible condition, $t(48) = 6.08$, $p < .001$ (one-sided), $d = 2.42$, 95% CI $[1.63, 3.20]$[6], as well as in the reversible condition, $t(48) = 4.11$, $p < .001$ (one-sided), $d = 1.23$, 95% CI $[0.70, 1.74]$. The ANOVA did not yield a significant interaction between causal reversibility and causal statement, $F(1, 48) = 1.93$, $p = .171$, $\eta_p^2 = .039$, which means that the way participants interpreted the two target causes did not differ significantly between the irreversible and the reversible causal structure condition. All in all, these results are similar to those observed in the original study.

The results of the novel part of the experiment in which participants rated the new trigger and maintainer statements for each of the two causes are summarized in Fig. 3. Participants in both reversibility conditions agreed that the first switch caused the turning on of the light, and at the same time disagreed that the second one did. Also, in the irreversible structure condition participants tended to disagree with the maintainer statements. In the reversible structure condition, the ratings for the trigger statements were very similar to

[6] Effect sizes were computed in R using the functions from the *MOTE* package. The code is included in the analysis script provided in the repository (see also Jané et al., 2024).

Figure 3

Participants' agreement ratings in the replication part of Experiment 1.



Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% confidence intervals. Jittered dots show participants' individual agreement ratings. b: Density plots showing the distribution of participants' agreement ratings.

those observed in the irreversible condition. What changed here, however, were the ratings for the maintainer statements. Participants tended to agree that the first cause was a maintainer of the effect. Importantly, this was also the case for the second cause, which shows that participants did perceive the second cause to be exerting an actual causal influence on the effect when the causal structure was reversible, as predicted by Ross and Woodward (2022). The density plots in Fig. 3b visualizing the distributions of the ratings show that this was true for a majority of participants in the reversible structure condition, while most participants disagreed when the structure was irreversible.

The descriptive pattern shown in Fig. 3 was analyzed with a mixed ANOVA with a Greenhouse-Geisser sphericity correction conducted with R's *afex* package (Singmann et al., 2022). As predicted, this analyses yielded a significant interaction effect between causal reversibility and causal statement, $F(2.92, 286.36) = 8.11$, $p < .001$, $\eta_p^2 = .076$[7]. This shows that how much participants agreed with the maintainer statements depended on whether the causal structure was irreversible or reversible; maintainer ratings were higher when the causal structure was reversible. Most importantly, participants' interpretation of the causal status of the second cause differed between the irreversible and reversible causal structure condition: Their agreement that the second cause is a maintainer of the effect was higher in the reversible condition than in the irreversible condition, as predicted by Ross and Woodward (2022). A planned directed contrast confirmed that this difference was significant, $t(98) = 4.32$, $p < .001$ (one-sided), $d = 0.86$, 95% CI $[0.45, 1.27]$.

A secondary finding was that the maintainer ratings for the first cause were also higher in the reversible structure condition than in the irreversible structure condition[8]. This result is plausible because the first cause (the first switch in the scenario) remains active after it turned on. It thus may be regarded to play a double role in this condition: initially, it triggers the effect but, because it remains active throughout the scenario, later also maintains it.

2.3 Conclusion

This study replicated the negative finding previously observed in the experiments reported by Henne et al. (2023). Looking at these results in isolation, one may conclude that reasoners do not think that the second cause in a preemption scenario makes an actual causal contribution to the target effect, be it under irreversible or reversible causal structures. However, in line with Ross and Woodward's (2022) hypothesis that the second cause is regarded as a maintainer (and not as a trigger of the target effect) in preemption scenarios that take place in reversible structures, the novel part of this study found that reasoners do think that the second cause makes an actual (maintaining) contribution to the effect.

A further observation was that the maintainer ratings, even under the reversible structure, remained lower than the trigger ratings for the first cause. One reason for this could be that a switch scenario is not optimal because switches are generally associated more with triggering than with maintaining. Another could be that some participants distributed their maintainer ratings between the two causes because they regarded them as equally contributing maintainers.

[7] The analysis also yielded a significant main effect of causal reversibility, $F(1, 98) = 23.28$, $p < .001$, $\eta_p^2 = .192$. This main effect was obtained because ratings were overall higher in the reversible condition. There was also a significant main effect of causal statement, $F(2.92, 286.36) = 98.83$, $p < .001$, $\eta_p^2 = .502$, which was driven by the differences between the two trigger statements (first caused vs. second caused). Planned directed contrasts conducted with R's *emmeans* package (Lenth et al., 2018) showed that the trigger statement about the first cause received higher ratings than that for the second cause in both the irreversible condition, $t(98) = 13.83$, $p < .001$ (one-sided), $d = 4.20$, 95% CI $[3.33, 5.08]$, and the reversible condition, $t(98) = 10.96$, $p < .001$ (one-sided), $d = 2.59$, 95% CI $[2.00, 3.17]$.

[8] $t(98) = 4.12$, $p < .001$, $d = 1.79$, 95% CI $[1.32, 2.25]$.

The remaining experiments aimed to generalize across different scenarios the central finding of the present study, which is that reasoners regard second causes as making an actual causal (maintaining) contribution to the target effect under reversible causal structures.

3 Experiment 2a

Experiment 2a tested a (fictitious) biological scenario about squids. A further difference from Experiment 1 is that this study used dynamic animations to convey the scenario and its structure to participants. A dynamic animation was also used to present the late preemption test situation. A further difference from Experiment 1 was the behavior of the first cause at the end of the test situation. In the scenario of Experiment 1, both light switches remained turned on. In the present study, the first cause turned off again a while after the second one had turned on. This was assumed to increase the maintainer ratings for the second cause and to lower the maintainer (but not the trigger) ratings for the first cause in the reversible structure condition.

The scenario described a newly discovered species of squids in which female squids can turn from brown to purple when approached by male squids. The scenario described a female squid and two males with the ability to make the female turn purple (and to stay purple in the reversible structure condition). An illustration of the stimulus material that was also presented to participants in the instructions is shown in Fig. 4. It was expected that this modified scenario would “work better” than the original switch scenario.

The materials of this study were pre-tested in a pilot study ($N = 93$) whose results can be accessed at the repository site. Experiment 2a’s pre-registration can be accessed here `https://osf.io/2x7yw` (Stephan, 2024b). A demo version of the study can be run at `https://simonstephan31.github.io/actual_cause_reverse/MainExperiments/Exp2a/experiment_files/task_local_demo/Exp_demo.html`.

3.1 Methods

3.1.1 Participants

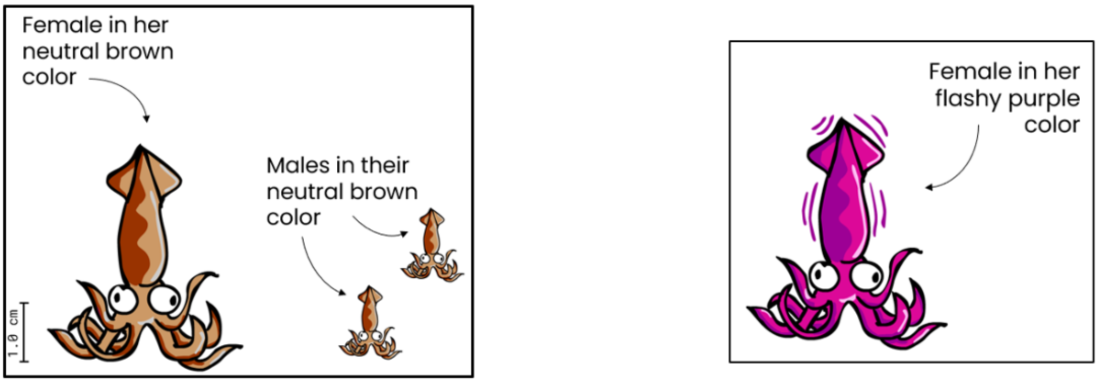
One hundred and eight participants took part in this experiment. The data of eight participants were excluded prior to any analyses because they failed to answer correctly at least one of the included control questions (as specified in the pre-registration). The final sample consisted of $N = 100$ participants ($M_{age} = 39.25$ years, $SD_{age} = 12.71$ years, age range 19 to 70 years) who were recruited via the online platform `www.prolific.co`. The inclusion and exclusion criteria were the same as in the previous experiment.

3.1.2 Sample size rationale

The sample size was based on an a priori power analysis. The effect size used in the planning was informed by what was observed in the pilot study. The basis of the sample size planning was the predicted interaction effect between *causal statement* and *causal reversibility* tested in a mixed ANOVA with R’s *afex* package. The goal was to reach at least 90% test power (for further details, see the pre-registration) for the detection of an effect size of $\eta_p^2 = .2$. The result of the analysis was a sample of $N = 60$. It was decided to test a larger sample of $N = 100$ ($n = 25$ in each between-participants condition, including counterbalancing conditions that will be described in more detail below).

Figure 4

Illustration of the scenario used in Experiment 2a.


Female in her neutral brown color
Males in their neutral brown color
1.0 cm
Female in her flashy purple color

3.1.3 Design, materials, and procedure

The study had the same mixed design as Experiment 1. Participants were alternately assigned to the between-participants conditions.

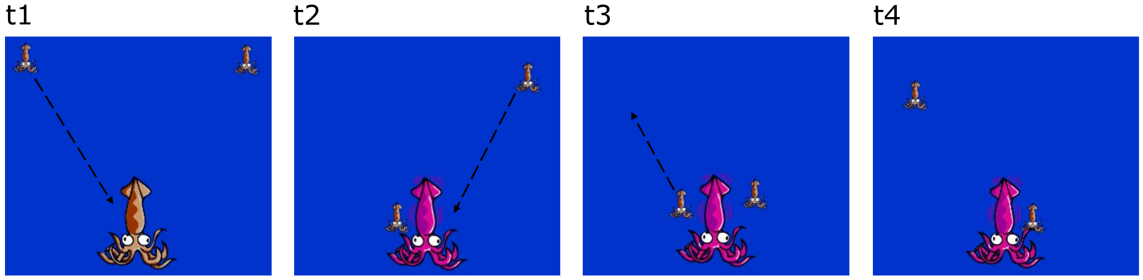
Before starting with the main part of the study, participants had to confirm that they were taking part via PC or laptop and that they were willing to pay attention. They also were shown a screen with a demo animation on which participants had to answer a multiple-choice question that probed whether they could see what happened in this animation. The purpose of this part was to ensure that participants were able to play the animations. It also allowed them to learn that they had to start each animations by clicking on a "Start" button displayed in the animation, and that each animation automatically returned to the starting frame once it had played to the end. The animations were created in Adobe Animate.

The main part of the experiment began with a general description of the scenario. It read:

> *Please read the following fictitious scenario:*
>
> Biologists have discovered a new species of squid. Females are large and males are small. Only females can change their color; they can turn purple. Illustrations are shown below.
>
> [The illustration was the one shown in Fig. 4]
>
> What causes a female to turn purple? A female squid turns purple if a male comes close to her.
>
> On the next screens, we will show you some animations allowing you to learn more about how male squids cause females to turn purple.
>
> *Please note*: You can watch all clips as often as you like, and we're always interested in your intuitive understanding.

Figure 5

Illustration of relevant sequences of the final test video of Experiment 2a.



Note. t_1*,* t_2*,* t_3*, and* t_4 *represent four points in time during the animation. Arrows visualize direction of movement and were added in this figure for illustrative purposes. They were not part of the animations.*

If you feel ready to start, please click "Continue" to proceed.

Participants then proceeded to a causal structure learning phase during which they were shown two learning animations, one for each of the two causes (the left and the right male squid) of the scenario. The spatial configuration of the squids at the beginning of each animation is illustrated in the first picture (t_1) of Fig. 5. These learning animations conveyed what happens to the female squid if only one male approaches her. Whether participants first saw the learning animation in which only the left squid interacts with the female or the animation in which only the right male interacts with the female was randomized between participants. In each clip, participants saw the respective male approach the female three times. As soon as a male came close to the female, she turned purple. Depending on the causal reversibility condition, the female either remained purple even when the male swim away from her again, or she immediately turned back to her natural brown color as soon as the male was far away enough. Participants had to answer a multiple-choice control question that was presented under each animation, which read: “Please select the option below that correctly describes what happened in the animation you’ve just seen”. Participants had to select one of the following two options: “When the left male approached the female she turned purple, but as soon as he swam away from her again, she turned back to her natural brown” versus “When the left male approached the female she turned purple and when he swam away from her again, she stayed purple”.

After these learning animations, participants proceeded to a novel screen where they had to answer a comprehension check question that probed their understanding of the scenario’s causal structure. It read: “Based on what you’ve learned from the two animations, which of the two options below correctly describes the color behavior of a female squid?” They had to select one of the following options: “For a female to turn purple, a male must come close to her. She stays purple only for as long as a male squid is close to her. As soon as a male swims away again, she immediately turns back to her natural brown” (correct answer in the reversible structure condition) versus “For a female to turn purple, a male must come close to her. Once this has happened, she remains purple permanently, even if a male squid swims away from her again” (correct answer in the irreversible structure condition). Only the data of participants were kept for analyses who passed this comprehension test.

During the test phase, the same animation was shown to participants in the irreversible and the reversible causal structure condition. An illustration of four crucial sequences of the final test animation is given in Fig. 4. In the beginning, the two male squids were in the upper corners of the screen. Then one of them began to swim towards the female (t_1) and when he arrived, she turned purple (t_2). Whether the first cause was the left or the right male was counterbalanced between participants; the sequence depicted in Fig. 5 shows the condition in which the left male served as first cause. While the first male was still close to the female (t_2), the other male began to approach the female. When he arrived (t_3), the other male returned to its initial position (t_4) and the animation ended shortly after that. The animation reset and participants could watch it again if they wanted.

The test question and the four actual causation statements were presented on the same screen below the animation. The question read: "How adequate is each of the following sentences to describe what happened in the clip you've just seen?". The four statements participants evaluated were presented in random order and read:

- The left male kept the female purple. [maintainer statement]
- The right male made the female turn purple. [trigger statement]
- The left male made the female turn purple. [trigger statement]
- The right male kept the female purple. [maintainer statement]

Participants provided their ratings for each statement on separate eleven-point rating scales, whose endpoints were labeled "Completely inadequate" and "Completely adequate".

On a novel screen, participants were asked a memory check question that was supposed to ensure that they were able to watch the test clip and did so thoroughly. They were asked to say which of the two male squids in the test animation swam to the female first (the left or the right one). Only the data of participants who answered correctly were kept for analyses.

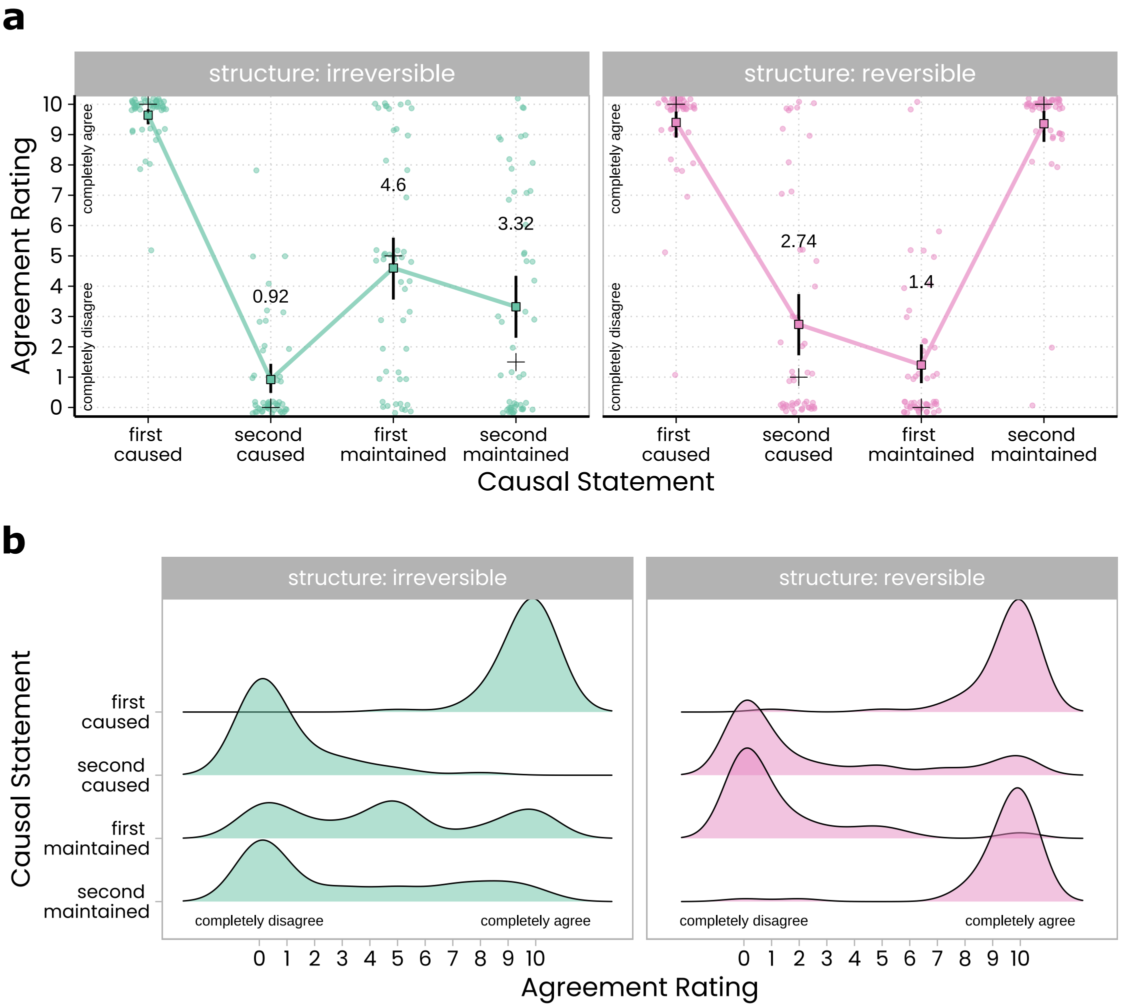
Participants then provided demographic data, could report any technical errors they might have encountered, and then finished the experiment on a short debriefing screen.

3.2 Results and discussion

Participants' ratings are summarized in Fig. 6. As can be seen there, the predicted pattern was observed. Irrespective of causal reversibility, participants strongly agreed that the first cause was a trigger of the effect ($M = 9.64$, 95% CI $[9.29, 9.99]$ in the irreversible condition vs. $M = 9.40$, 95% CI $[9.05, 9.75]$ in the reversible condition), while they strongly disagreed that the second cause was a trigger of the effect ($M = 0.92$, 95% CI $[0.11, 1.73]$ in the irreversible condition vs. $M = 2.74$, 95% CI $[1.93, 3.55]$ in the reversible condition). By contrast, the ratings for the maintainer statements were very different depending on causal reversibility. As predicted, the maintainer ratings for the second cause were much higher when the structure was reversible ($M = 9.36$, 95% CI $[8.55, 10.17]$) than when it was irreversible ($M = 3.32$, 95% CI $[2.51, 4.13]$). The density plot in Fig. 6b shows that most participants in the reversible structure condition agreed strongly with the maintainer

Figure 6

Participants' agreement ratings in Experiment 2a.



Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% confidence intervals. Jittered dots show participants' individual agreement ratings. b: Density plots showing the distribution of participants' agreement ratings.

statement about the second cause. Also as expected, maintainer ratings for the first cause were lower this time in the reversible condition (compared to what was observed in Experiment 1). This is a plausible result because in the present experiment the first cause disappeared again shortly after it had triggered the effect and the second cause had turned on. Notably, maintainer ratings for the second cause this time were almost as high as the trigger ratings for the first cause. This indicates that, unlike in classic late preemption scenarios in irreversible structures, participants had a clear intuition that the second cause had an actual causal impact on the effect.

The pattern shown in in Fig. 6 was tested by a mixed ANOVA with Greenhouse-Geisser sphericity correction conducted with R's *afex* package (Singmann et al., 2022) and by planned (directed) contrasts conducted with R's *emmeans* package (Lenth et al., 2018). The results corroborated the predictions. The ANOVA yielded a significant interaction effect between causal reversibility and causal statement, $F(2.67, 261.83) = 52.73$, $p < .001$, $\eta_p^2 = .350$[9]. As predicted, a planned (directed) contrast confirmed that participants agreed significantly more with the maintainer statement for the second cause when they had learned that target effect was reversible, $t(98) = 10.44$, $p < .001$, $d = 2.09$, 95% CI $[1.60, 2.57]$.

3.3 Conclusion

The results of this study replicate in a new test scenario what was previously found in the novel part of Experiment 1: causal reversibility has an effect on reasoners' perception of the second cause in a preemption-like sequence of events. If they know that the effect reverses if its causes reverse, they tend to perceive the second cause to make an actual, maintaining, causal contribution to the effect.

Another aspect of this experiment that may be stressed is that participants in the reversible structure condition actually never saw the effect disappear again in the final test animation. The female squid turned purple when the first male came close to her and she then remained purple even when the first male disappeared again, because the second male was already in in place at that moment. In fact, subjects in the reversible structure condition saw the exact same final animation as participants in the irreversible structure condition. Therefore, the higher maintainer ratings for the second cause that were measured in the reversible structure condition must have resulted from a counterfactual reasoning process on the site of the participants: what led participants to give high maintainer ratings for the second cause was that they knew that the effect would have disappeared again if the second cause had disappeared again. In sum, the study provides further evidence for the hypothesis of Ross and Woodward (2022).

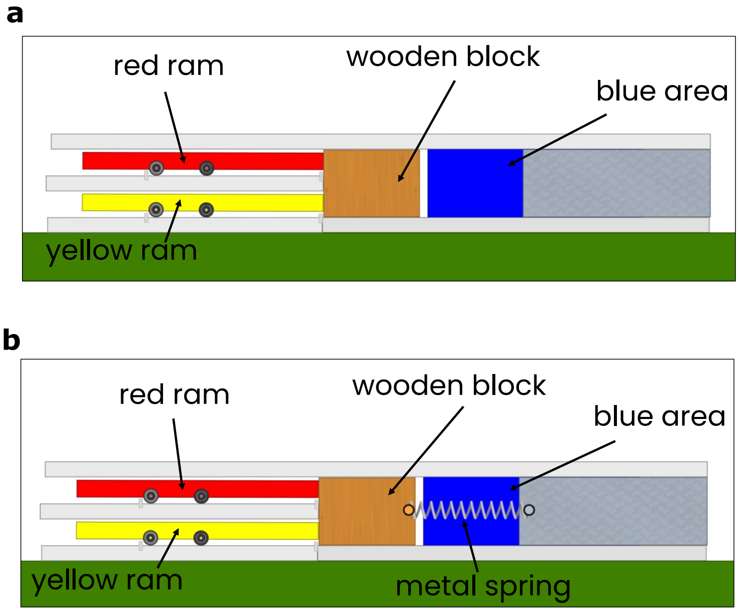
4 Experiment 2b

Experiment 2a replicated the central findings of Experiment 1a in a novel scenario. However, although Experiment 2a's scenario was from a completely different domain (biology) than the light switch scenario of Experiment 1a, the "behavior" of the effect was very similar: the female squid changing color in this scenario was very similar to the purple light turning on in the switch scenario. Experiment 2b thus sought to generalize the findings of the previous studies further by testing yet another scenario. This time, the scenario was a mechanical scenario about rams pushing a wooden block into a target area. An illustration of the scenario is shown in Fig. 7.

[9] The analysis also yielded a significant main effect of causal reversibility, $F(1, 98) = 19.03$, $p < .001$, $\eta_p^2 = .163$. This main effect was obtained because ratings were overall higher in the reversible condition. There was also a significant main effect of causal statement, $F(2.67, 261.83) = 168.69$, $p < .001$, $\eta_p^2 = .633$, which was driven by the differences between the four different causal statements. Planned directed contrasts showed that the trigger statement about the first cause received higher ratings than that about the second cause in both the irreversible condition, $t(98) = 17.72$, $p < .001$ (one-sided), $d = 6.57$, 95% CI $[5.24, 7.87]$, and the reversible condition, $t(98) = 13.53$, $p < .001$ (one-sided), $d = 2.43$, 95% CI $[1.87, 2.98]$.

Figure 7

Illustration of the mechanical scenario of Experiment 2b.



Note. a: illustration of the irreversible version of the scenario. b: reversible version of the scenario.

In this novel scenario, two rams on wheels (a red ram and a yellow ram) were placed on different tracks on top of each other. They touched a wooden block that each ram on its own could push to the right into a blue target area. The difference between the irreversible (Fig. 7a) and the reversible structure condition (Fig. 7b) was the presence (in the reversible structure condition) or absence of a metal spring (in the irreversible structure condition) that connected the wooden block with the wall behind the blue area. The chosen values for the spring's target length and for the spring constant that were specified in the physics simulator used to create the scenario ensured that the spring keeps the wooden block out of the blue area when no ram is pushing it. In the irreversible case, where no metal spring was present, the friction parameters of the surfaces were set up such that the wooden block remains in the blue area if it gets pushed into it.

As in the previous studies, the prediction was that participants would agree that the second cause in a preemption-like sequence exerts a maintaining influence on the effect in a reversible causal structure but not in an irreversible structure. If the causal structure is irreversible, participants were expected to think that the first cause preempts the second. Experiment 2b's pre-registration can be accessed here `https://osf.io/edhnb` (Stephan, 2024d). A demo version of the experiment can be run at `https://simonstephan31.github.io/actual_cause_reverse/MainExperiments/Exp2b/experiment_files/task_local_demo/Exp_demo.html`.

4.1 Methods

4.1.1 Participants

One hundred and eleven participants took part in this experiment. The data of eleven participants were excluded prior to any analyses because they failed to answer a control question correctly (as specified in the pre-registration). The final sample thus consisted of $N = 100$ participants ($M_{age} = 41.83$ years, $SD_{age} = 13.30$ years, age range 19 to 76 years) recruited via the online platform `www.prolific.co` participated in this online study and provided complete data. The inclusion and exclusion criteria used for the recruitment of participants from prolific participant pool were the same as in the previous experiment.

4.1.2 Sample size rational

The sample size was based on an a priori power analysis. The effect size used in the planning was informed by what was observed in the pilot study. The basis of the sample size planning was the predicted interaction effect between *causal statement* and *causal reversibility* tested in a mixed ANOVA with R's *afex* package. The goal was to reach at least 90% test power (for further details, see the pre-registration) for the detection of an effect size of $\eta_p^2 = .2$. The result of the analysis was a sample of $N = 60$. It was decided to realize a larger sample of $N = 100$ ($n = 25$ in each between-participants condition, including counterbalancing conditions that will be described in more detail below).

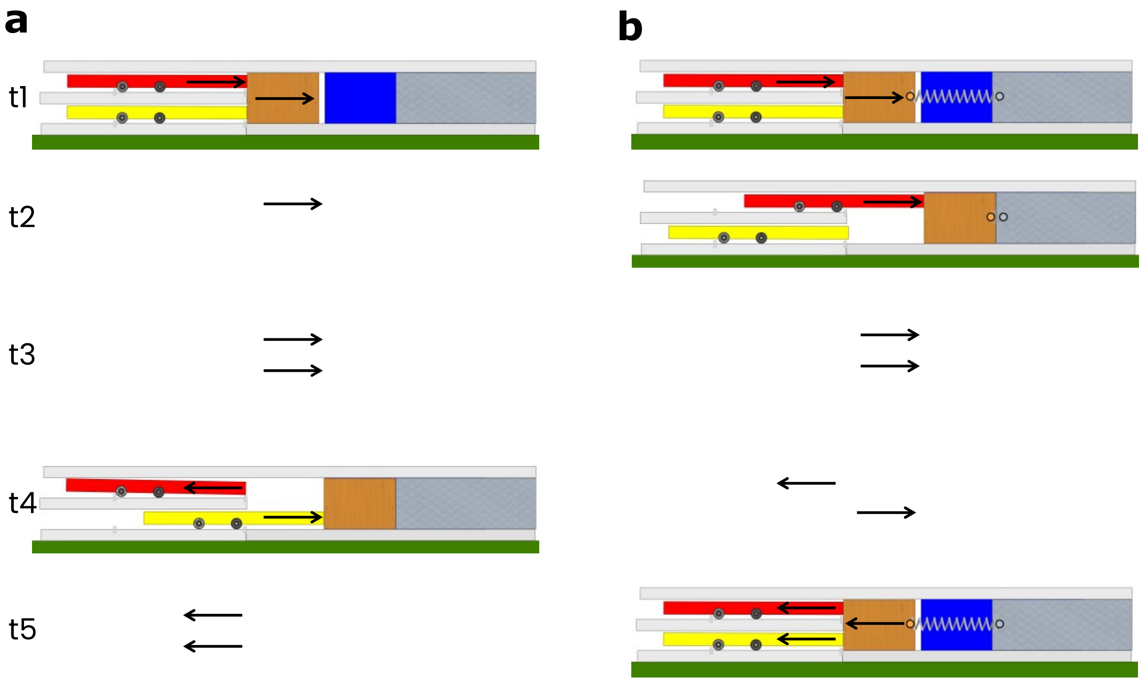
4.1.3 Design, materials, and procedure

The experiment had the same study design as Experiment 2a and an overall very similar procedure. The only relevant difference was the experimental scenario. On the first screen of the scenario introduction, participants learned about the mechanical arrangement of the two rams, the wooden block, and the blue area. Depending on the causal reversibility condition, participants were shown an illustration of the "machine" that looked like the one in Fig. 7a (in the irreversible structure condition) or like Fig. 7b (in the irreversible structure condition). They then were shown two short learning clips, one for each of the two rams. These video clips allowed them to learn the causal reversibility of the scenario because the clips showed what happens to the wooden block if it gets pushed by a ram and also what happens to it if a ram returns to its initial position. The fictitious machine was created and simulated in the physics simulator Algodoo (`http://www.algodoo.com/`), and the video clips participants were shown were created using the screen recorder coming with Microsoft's "snipping tool". For each of the two videos shown in the learning phase, participants had to answer a multiple-choice question that asked them to select the option that correctly described what happened in the learning clip they just saw. The two options were: "When the yellow [red] ram returned to its initial position, the wooden block remained in the blue area" (correct option in the irreversible structure condition) and "When the yellow [red] ram returned to its initial position, the wooden block also went back (left the blue area again)" (correct option in the reversible structure condition). Only the data from participants were kept for analyses who selected the correct option in each of the two learning videos.

The final test video showed a situation that instantiated the temporal pattern of a late preemption scenario: one of the two rams began to move before the other and reached the wooden block first. Whether the yellow or the red ram moved first was counterbalanced

Figure 8

Illustration of relevant sequences of the final test videos of Experiment 2b.



Note. The five sequences (t_n) shown in this figure are from the version of the test videos in which the red ram moved first and the yellow ram second. a: sequences of the test video shown in the irreversible causal structure condition. b: sequences of the test video shown in the reversible causal structure condition.

between participants. An illustration of relevant sequences of the test clip presented in the different conditions is shown in Fig. 8. Either the red or the yellow ram started to move and push against the wooden block first. In Fig. 8, the first ram (the first cause) starting to move is the red one (t_1). It pushed against the wooden block, which then moved into the blue area (t_2). Then the yellow ram began to move to the right until it also made contact with the wooden block (t_3). After a short moment, the first ram began to return to its initial position (t_4) while the second ram remained in contact with the wooden block. It then also returned to its initial position. What happened to the wooden block at this point depended on whether the scenario was causally reversible or not (t_5). Thus, the test situation at this point deviated from what participants saw in the previous experiments. There, at least one (the second cause in Experiment 2a) or both causes remained present (or "on") in the test situation after they had become active. In the present test case, by seeing whether the effect actually reversed or not when the second cause disappeared again, participants had an additional diagnostic cue for the causal reversibility of the scenario.

The test question and the four causal statements were displayed on the same screen below the video clip. The test query read: "How adequate is each of the following sentences to describe what happened in the clip you've just seen?" The four causal statements were presented in random order. Participants evaluated them on the same eleven point rating scale that was used in Experiment 2a. The statements were:

- The yellow ram made the wooden block stay in the blue area for a while. [maintainer statement]
- The red ram caused the wooden block to go into the blue area. [trigger statement]
- The red ram made the wooden block stay in the blue area for a while. [maintainer statement]
- The yellow ram caused the wooden block to go into the blue area. [trigger statement]

After the test screen, participants had to answer a control query asking them to say which of the two rams had moved first in the test video (red vs. yellow). Only the data of participants who answered correctly were kept for analyses. The remainder of the experiment was like in the previous studies.

4.2 Results and discussion

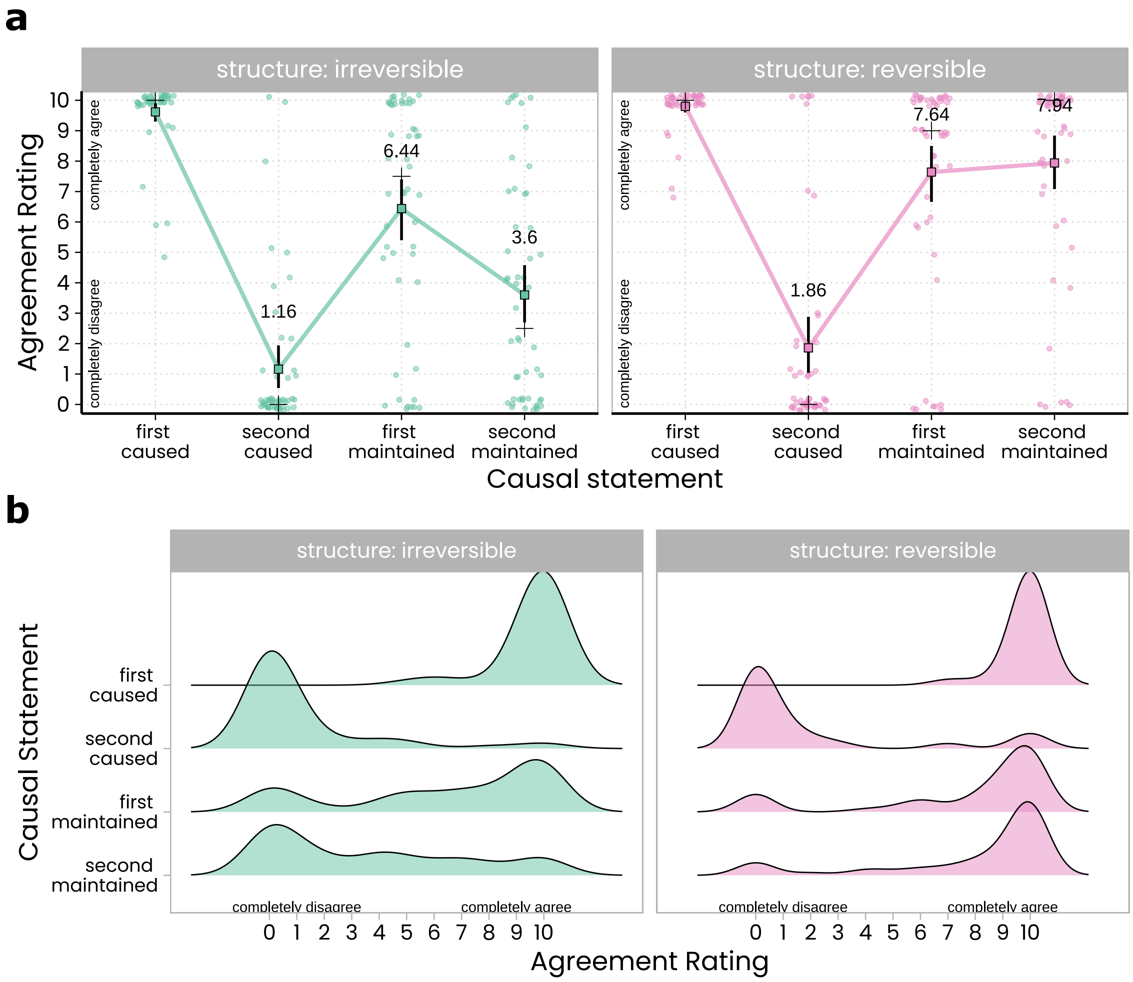
Participants' ratings are summarized in Fig. 6. It shows that the main finding from the previous experiment could be replicated also with the new mechanical scenario. As before, the first cause was considered to be a trigger of the effect irrespective of causal reversibility ($M = 9.62$, 95% CI $[9.36, 9.90]$ in the irreversible structure condition and $M = 9.80$, 95% CI $[9.54, 10.06]$ in the reversible structure condition). At the same time, participants expressed that the second cause is not a trigger of the effect, independent of causal reversibility ($M = 1.16$, 95% CI $[0.33, 1.99]$ in the irreversible structure condition and $M = 1.86$, 95% CI $[1.03, 2.69]$ in the reversible structure condition). What was influenced by causal reversibility was how participants reacted to the maintainer statements, especially when they were about the second cause. As predicted by the Ross-Woodward hypothesis, participants agreed that the second cause exerted a maintaining influence on the effect, but only when the causal structure was reversible ($M = 7.64$, 95% CI $[6.63, 8.65]$ in the reversible structure condition and $M = 3.60$, 95% CI $[2.64, 4.56]$ in the irreversible structure condition). The density plot in Fig. 9b shows that a majority of participants strongly agreed that the second cause was a maintainer.

The statistical analyses corroborated the descriptive pattern shown in Fig.9. A mixed ANOVA with Greenhouse-Geisser sphericity correction conducted with R's *afex* package (Singmann et al., 2022) yielded the predicted interaction between causal reversibility and causal statement, $F(2.67, 261.70) = 10.65$, $p < .001$, $\eta_p^2 = .10$[10]. As can be seen in Fig. 9 this interaction was mostly driven by the difference in the maintainer ratings for the second cause. A planned directed contrast confirmed that maintainer ratings were

[10] The analysis also yielded a significant main effect of causal reversibility, $F(1, 98) = 26.61$, $p < .001$, $\eta_p^2 = .214$. This main effect was obtained because ratings were overall higher in the reversible condition. There was also a significant main effect of causal statement, $F(2.67, 261.70) = 142.29$, $p < .001$, $\eta_p^2 = .592$, which was driven by the differences between the four different causal statements. Planned directed contrasts showed that the trigger statement about the first cause received higher ratings than that for the second cause in both the irreversible condition, $t(98) = 17.96$, $p < .001$ (one-sided), $d = 4.54$, 95% CI $[3.60, 5.46]$, and the reversible condition, $t(98) = 16.86$, $p < .001$ (one-sided), $d = 3.45$, 95% CI $[2.71, 4.18]$.

Figure 9

Participants' agreement ratings in Experiment 2b.



Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% confidence intervals. Jittered dots show participants' individual agreement ratings. b: Density plots showing the distribution of participants' agreement ratings.

significantly higher in the reversible than in the irreversible causal structure condition, $t(98) = 6.33$, $p < .001$ (one-sided), $d = 1.27$, 95% CI $[0.83, 1.69]$.

If the maintainer ratings for the first cause in the reversible condition are compared with the corresponding ones from Experiments 1 and 2a, it can be seen that participants this time gave higher maintainer ratings for the first cause in the reversible condition, like in Experiment 1. This is explained by the fact that in this scenario there was a moment at which both rams remained in contact with the wooden block (see t3 in Fig. 8). Thus, similar to Experiment 1 and unlike in Experiment 2a, there was a moment at which both causes remained active together. It thus appears plausible to say that the first cause not only triggered the effect, but also maintained it (at least for a certain period). Surprisingly, the first cause in the irreversible structure condition also received high maintainer ratings.

4.3 Conclusion

Like the previous ones, this experiment shows that second causes in a preemption-like sequence are perceived as exerting an actual causal influence on a target effect; what is necessary for this to happen is that the causal structure is reversible. In irreversible structures, by contrast, causes happening second tend to be regarded as non-causal. Here, the first cause is perceived to preempt the second cause in exerting an actual influence on the effect. The study corroborates Ross and Woodward's (2022) hypothesis.

5 General Discussion

Most philosophical theories of actual causation are based on the premise that causal relations hold between *events* – especially those theories belonging to the class of so-called counterfactual dependency theories (see Beebee, Hitchcock, & Menzies, 2009; Woodward, 2021, for an overviews on this and other classes of causality theories). Most of these theories analyze scenarios in which the focus is on the moments at which events happen, that is, their *onsets*. Examples of events starring in famous philosophical thought experiments on actual causation are victims getting poisoned by assassins (see, e.g., Halpern & Hitchcock, 2015; Hitchcock, 2007), bottles getting destroyed by rocks (see, e.g., Hall, 2004), and forests starting to burn after lightning strikes (see, e.g., Halpern & Pearl, 2005a, 2005b), to name just a few. This focus on event causation led to the neglect of other aspects of causal relations (but see also Woodward, 2006, 2010, who analyzes the aspect of causal stability). One such aspect of causal relations is their "reversibility" (Ross & Woodward, 2022), which was the focus of the present paper. In recent theoretical work, Ross and Woodward (2022) began to look at causal reversibility and its possible consequences. They did so by focusing on a specific kind of situation known as *late preemption*. The reason why this kind of scenario is interesting is that it has been taken to reveal a very stable causal intuition: In all scenarios of causal preemption that have been exercised in philosophical debates, or subjected to empirical psychological tests, it has been taken for granted that the cause that happens first is the actual cause of the target effect, while the one that happens second is not a cause at all (in this specific situation). It seemed that all the ingredients it takes to create this impression of causal preemption is a situation in which two perfectly reliable general causes of a target effect are instantiated at different points in time (but see also Stephan et al., 2020, who demonstrate that reasoners' knowledge about the causal latency of the first and second cause also matters). Ross and Woodward (2022) speculated that a crucial factor remained unconsidered, and that our seemingly rock solid intuition of causal preemption in such a sequence of events might change if we changed only one aspect about the situation: the reversibility of the causal structure in which the events take place. Causal reversibility (or rather irreversibility) be, so to say, another relevant ingredient that had implicitly been present in the classic preemption scenarios. According to the Ross-Woodward hypothesis, second causes in a preemption-like sequence of events can be perceived to exert an actual causal influence on the target effect if the causal structure is reversible.

Henne et al. (2023) were the first who sought to put the hypothesis to test, to see if lay people's actual causation judgments are in line with it. Their results were negative, which led them to conclude: “We failed to find evidence that reversibility affects causal judgments in cases of late preemption. Instead, we found that people judge that the earlier preempting event is more causal than the later preempted alternative event – regardless of reversibility – which is consistent with previous work on late preemption (Henne et al., 2021; Lombrozo, 2010; Walsh & Sloman, 2011)” (p. 15).

The central hypothesis of the present paper was that the theoretical analysis by Ross and Woodward (2022) predicts a psychologically real phenomenon and that Henne et al. (2023) missed to observe it in their studies because they did not use test questions that adequately captured the nature of the second cause of preemption-like sequences in a reversible causal structure. A crucial observation made by Ross and Woodward (2022) was that second causes in this kind of situation seem to be *maintainers* of the effect. By presenting causal statements to their participants that might be interpreted as claims referring to the *trigger(s)* of the target effect, Henne et al. (2023) might have missed to uncover the predicted effect. In light of the results of the present experiments, it seems that this was indeed the case. The problem that test questions are not always unambiguous with respect to the kind of cognitive process or judgment they aim to assess is not new in the causal reasoning literature. For example, a similar observation was made by Griffiths and Tenenbaum (2005) when they showed that (many) participants interpreted questions intended to asses knowledge about the strength (or power) of causal relations (Cheng, 1997) as questions asking about causal structure. Similarly, Cheng and Novick (2005) showed that studies that seemingly provided evidence against the power PC theory (Cheng, 1997) – a computational theory about causal strength learning – actually asked participants causal attribution instead of causal strength queries.

Some aspects about the present set of studies might be criticized. One aspect is that the studies manipulated the causal test statements within participant. The simultaneous presentation of all the different statements may have encouraged participants to think about the differences between triggers and maintainers. While this aspect cannot explain the predicted interaction between causal structure (which was manipulated between participants) and causal test statement, it is still an interesting question how similar the results would look like if the test statements were manipulated between participants. For this reason, a supplementary study ($N = 240$) was conducted, whose results are reported and whose materials were made available (including a demo of the study) on the repository site (see `https://simonstephan31.github.io/actual_cause_reverse/expSup_mat.html#53_Results_and_discussion`). In this study, both causal structure and causal test statements were manipulated between participants. The experimental scenario was the squid scenario from Experiment 2a. Except for the test phase, where participants this time evaluated only a single test statement, the procedure was identical to the one of Experiment 2a. The results of this study closely replicated the findings of the main experiments reported in this paper.

Another aspect that might be criticized about the present experiments is that the test statements could have been phrased in a more parallel way. In all the experiments, the formulations of the maintainer statements were of the form “A kept B C-ing” (e.g., “At 4:00PM, the blue switch being on is keeping the purple light burning” or “The left [right] male kept the female purple”), while the trigger statements used a less direct construction of the form “A caused/made B to C” (e.g., “The red ram caused the wooden block to go into the blue area" or "The right male made the female turn purple"). Interestingly, in a recent study Rose, Sievers, and Nichols (2021) have shown that whether causal statements are formulated in the indirect form of "A caused B to C" or in a direct form using causatives (e.g., "Antonia caused the vase to break" vs. "Antonia broke the vase") can have an influence on reasoners' causal judgments[11]. While this difference cannot explain the predicted effect of the Ross-Woodward hypothesis in the present paper, it still seems desirable to have more parallel formulations. One possibility is to use trigger statements that avoid the indirect construction "caused to" by using a causative. For example, the trigger statements in Experiment 2a could have been "The left [right] male *turned* the female purple" rather than "The right male made the female turn purple". To address this aspect, the supplementary study reported on the repository site used more parallelized test statements; both the trigger and maintainer statements were formulated in the direct form. As has been mentioned above, the study still replicated the findings of the other experiments in this paper and still confirmed the Ross-Woodward hypothesis.

The present study's corroboration of the Ross-Woodward hypothesis is relevant for future research. First of all, the finding that lay people perceive second causes in preemption-like sequences as making an actual (maintaining) causal contribution to the effect if the causal structure is reversible might have implications for theories of actual causation. Most counterfactual theories of actual causation are triggering accounts of actual causation, while the concept of maintaining has widely been neglected. This may seem surprising given that maintaining is prevalent in our world. Some examples of maintaining we frequently experience were given earlier in this article. Investigating maintaining relations seems interesting both from a theoretical and a psychological perspective because in situations in which a maintaining relation is active, nothing visible seems to be happening. This is because maintaining seems to be a relation between standing states rather than a relation between events. Thus, there is not only the question of how maintaining relations can be captured theoretically but also of how people succeed in detecting them.

Understanding how people infer and represent maintaining relations is also relevant for research studying the language of causation (Neeleman & van de Koot, 2012) and their underlying mental representation. So far, psychological theories and experiments in that area have addressed causal concepts like "enable" or "allow" and what distinguishes them from the meaning of "cause" (Beller, Bennett, & Gerstenberg, 2020; Cao, Geiger, Kreiss, Icard, & Gerstenberg, 2023; Cheng & Novick, 1991, 1992; Sloman, Barbey, & Hotaling, 2009; Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010; Wolff & Song, 2003), but studies that (directly) address how people learn and think about causal maintaining are still largely absent. An exception is a recent paper by Zhou, Smith, Tenenbaum, and Gerstenberg (2023), who propose a counterfactual simulation model of *physical support*. Their model is able to capture reasoners' judgments in scenarios in which building blocks of a tower (e.g., the wooden blocks of a jenga tower) are either actually or only hypothetically removed. The model accurately predicts the degree to which participants consider certain blocks to be crucial for a towers stability. This can be conceptualized as a judgment about maintaining, since a block perceived to be crucial for the stability of a tower could be said to "maintain" its stability. The application of their model is so far restricted to relatively simple physical

[11] I would like to thank Paul Henne for pointing this out.

scenes, however.

The model by Zhou et al. (2023) is a counterfactual dependency model. Another interesting question is how maintaining might be modeled by different classes of models about causal semantics, like force dynamics models (Talmy, 1988; Wolff, 2007; Wolff et al., 2010; Wolff & Song, 2003; Wolff & Thorstad, 2017), covariation accounts (e.g., Cheng & Novick, 1991, 1992), and causal Bayes nets (e.g., Sloman et al., 2009). One line of future studies, for example, could investigate to which extent the force dynamics model's conceptualization of "prevent" may already implicitly incorporate the meaning of "maintain". Analyzing the force configuration between entities called "afector" and "patient", the model is able, for example, to capture scenes that a human speaker might describe by saying "A prevents B from C-ing". At first glance, it seems that maintaining relations can (often) be described as preventing relations of this form. For instance, the sentence "the male squid kept the female purple" could be rephrased as "the male squid prevented the female from turning brown again". It might be interesting to study to which extent reasoners do indeed regard both as being equivalent. In this context, it may also be interesting to study which causal statement constructions people use spontaneously to pragmatically convey a maintaining relation. One hypothesis is that reasoners are more inclined to endorse maintainer statements of the form "A kept B C-ing" because this kind of statement directly expresses the core notion of maintaining, which is the persistence of the current state of an effect, whereas a prevent-type formulation refers to a *change* of the effect (i.e., an event) that is actually not happening.

The present paper, together with Ross and Woodward's (2022) theoretical analysis, suggests that a crucial factor allowing people to infer a maintaining relation is knowledge about the reversibility of the causal structure[12]. This paper did not look beyond reversibility in preemption-like sequences, however, which is why open questions remain. For example, looking back at the results of Experiment 2b, it was found that the first cause received high maintainer ratings in the irreversible structure condition (to a lesser degree this also happened in Experiment 2a). Also, the maintainer ratings for the first cause of the irreversible structure condition were always found to be higher than the trigger ratings for the second cause. If this is a reliable finding, the question is why people may think that a cause maintains an effect even though they are aware that the effect would not disappear if the cause was removed. It seems that this judgment pattern cannot be explained by counterfactual accounts, which makes it an interesting target for future investigations. Unless this pattern is the result of mere stochastic noise, an interesting question is what specific aspects about the scenarios used in Experiments 2a and 2b prompt this intuition.

In future studies, it would also be interesting to go beyond actual causation judgments and look at the role of causal reversibility in the induction of general causal relations, or the selection of causal interventions. A widely accepted view is that general causal relations are induced based on observable covariations (Cheng, 1997; Cheng & Novick, 1990; Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014; Novick & Cheng, 2004), and some studies have begun to look more closely at how covariation interacts with temporal information in reasoners' causal learning (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Gong, Gerstenberg, Mayrhofer, & Bramley, 2023; Greville & Buehner, 2010;

[12] Such knowledge might also be at play in the mental jenga task of Zhou et al. (2023), as participants might be very much aware that a tower that was built is able to collapse again.

Hagmayer & Waldmann, 2002). These latter studies find that a crucial temporal factor is causal latency, which can be defined as the time it takes a cause to produce an effect (see also Stephan et al., 2020; Stephan & Waldmann, 2022). Future studies could look at how temporal features of reversible causal structures figure into causal learning and reasoning. For example, in addition to the latency and the strength of a cause, another relevant aspect about a causal relation that would be useful for a reasoner to learn is the "persistence" with which a cause maintains its effect. For example, if an intervention reliably generates a desired effect only for a short period of time, it might be better to opt for an alternative intervention that may be less reliable but have a longer-lasting effect.

Finally, the findings in this paper may be relevant not only for the analysis of how people causally explain singular effects in specific situations (i.e., actual causation judgments), but also for the (philosophical and psychological) study of explanation more generally. For example, an often-discussed concept in philosophical studies on explanation that can be related to the concept of maintaining is that of "constraints" or "structural factors"[13] (see, e.g., Dretske, 1988; Haslanger, 2016; Ross, 2023). When an explanation of a target system (e.g., why does a toy boat on a river take specific path?) involves constraints (e.g., the river banks constraining the path the toy boat can take; see Ross, 2023), these constraints seem to be conceptually similar to maintainers that keep a system in a particular state (or a particular range of states). An analogy used by (Ross, 2023, p. 5) to illustrate the concept of constraint is that of "a building's frame and an organism's skeletal structure, which are fixed, physical scaffolds that limit various outcomes of these systems". The present studies suggest that the concept of constraining/maintaining is not a concept that can only be found in philosophical armchair analyses. Rather, it seems to be a readily available tool in lay people's (causal) explanations.

6 Conclusion

Causal reversibility can drastically change how we think about the causal status of events. For example, second events in preemption-like sequences that seem to be genuine non-causes if the scenario has an irreversible causal structure can turn into actual causes if the causal structure is reversible. They are conceptualized as maintainers in this case. Future studies should continue to investigate the concept of maintaining and of causal reversibility.

7 Acknowledgments

The author would like to thank Sarah Placì for helpful comments and Paul Henne as well as the remaining, anonymous, reviewers for the constructive evaluation of this work. This research was funded by the German Research Foundation (DFG) in the form of a Reinhart Koselleck project (WA 621/25-1).

8 References

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The Oxford handbook of causation.* Oxford University Press.

[13] I would like to thank an anonymous reviewer for pointing this out.

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In S. Denision, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (p. 3133-3139). Cognitive Science Society.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880—1910.

Cao, A., Geiger, A., Kreiss, E., Icard, T., & Gerstenberg, T. (2023). A semantics for causing, enabling, and preventing verbs using structural causal models. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Meeting of the Cognitive Science Society* (pp. 2947–2954). Cognitive Science Society.

Chang, W. (2009). Connecting counterfactual and physical causation. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1983–1987). Cognitive Science Society.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545–567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*(1–2), 83–120.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382.

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*(3), 694–706.

Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 201–215). New York: Oxford University Press.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*, 5351.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes.* MIT press.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*, 936–975.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, 101542.

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–771.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*, 1128–1137.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 225–276). Cambridge: The MIT Press.

Halpern, J. Y. (2016). *Actual causality.* MIT Press.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457.

Halpern, J. Y., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Halpern, J. Y., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, *56*(4), 889–911.

Haslanger, S. (2016). What is a (social) structural explanation? *Philosophical Studies*, *173*, 113–130.

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

Henne, P., Perez, K., & McCracken, C. (2023). *No evidence that reversibility affects causal judgments in late-preemption cases.* Retrieved from `https://doi.org/10.31219/osf.io/ky9xn` (Preprint)

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*, 273–299.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, *116*(4), 495–532.

Hitchcock, C. (2009). Causal modelling. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 299–314). New York: Oxford University Press.

Jané, M. B., Xiao, Q., Yeung, S. K., Ben-Shachar, M. S., Caldwell, A. R., Cousineau, D., ... Feldman, G. (2024). *Guide to effect sizes and confidence intervals.* Retrieved from `https://matthewbjane.quarto.pub/effect-size-and-confidence-intervals-guide/`

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Package "emmeans"' [Computer software manual]. Retrieved from `https://cran.r-project.org/web/packages/emmeans/index.html` (R package version 4.0-3)

Lewis, D. (1973). Causation. *The journal of philosophy*, 556–567.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*, 182–197.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277–301.

Neeleman, A., & van de Koot, H. (2012). The linguistic expression of causation. In M. Everaert, M. Marelj, & T. Siloni (Eds.), *The theta system – argument structure at the interface* (pp. 20–51). New York: Oxford University Press.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–485.

Paul, L. A. (2009). Counterfactual theories. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 158–184). New York: Oxford University Press.

Paul, L. A., & Hall, E. J. (2013). *Causation: A user's guide.* New York: Oxford University Press.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rose, D., & Danks, D. (2012). Causation: Empirical trends and future directions. *Philosophy Compass*, *7*(9), 643–653.

Rose, D., Sievers, E., & Nichols, S. (2021). Cause and burn. *Cognition*, *207*, 104517. Retrieved from `https://www.sciencedirect.com/science/article/pii/S001002772030336X` doi: https://doi.org/10.1016/j.cognition.2020.104517

Ross, L. N. (2023). The explanatory nature of constraints: Law-based, mathematical, and causal. *Synthese*, *202*(56).

Ross, L. N., & Woodward, J. (2022). Irreversible (one-hit) and reversible (sustaining) causation. *Philosophy of Science*, *89*(5), 889–898.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=afex` (R package version 1.1-1)

Sloman, S., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, *33*, 21–50.

Stephan, S. (2024a). *Actual cause judgments in irreversible and reversible causal structures.* Retrieved from `https://doi.org/10.17605/OSF.IO/DNBF6`

Stephan, S. (2024b). *Actual cause judgments in irreversible and reversible causal structures 02.* Retrieved from `https://doi.org/10.17605/OSF.IO/2X7YW`

Stephan, S. (2024c). *Actual cause judgments in irreversible and reversible causal structures 03.* Retrieved from `https://doi.org/10.17605/OSF.IO/TWY6C`

Stephan, S. (2024d). *Actual cause judgments in irreversible and reversible causal structures 04.* Retrieved from `https://doi.org/10.17605/OSF.IO/EDHNB`

Stephan, S. (2024e). *Reasoning about actual causation in reversible and irreversible causal structures.* Retrieved from `https://simonstephan31.github.io/actual_cause_reverse/`

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation – a computational model. *Cognitive Science*, *44*(7), e12871.

Stephan, S., & Waldmann, M. R. (2017). Preemption in singular causation judgments: A computational model. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1126–1131). Austin, TX: Cognitive Science Society.

Stephan, S., & Waldmann, M. R. (2022). The role of mechanism information in singular causation judgments. *Cognition*, *218*, 104924.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning.* New York: Oxford University Press.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, *26*, 21–52.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, *47*, 276–332.

Wolff, P., & Thorstad, R. (2017). Force dynamics. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 147–168). New York: Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*, 1–50.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*, 287–318.

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology.* Oxford University Press.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*, *152*, 2237–2269.