Probing counterfactual thought without counterfactual language

David Rose[1], Siying Zhang[2], Sophie Bridgers[1], Hyowon Gweon[1], and Tobias Gerstenberg[1]

[1]Stanford University, Department of Psychology

[2]University of Washington, Department of Psychology

Abstract

Counterfactual thinking—thinking about how things could have gone differently—is a hallmark of human intelligence. It plays a central role in determining what causes what, experiencing regret, engaging in pretense, in understanding rules and mental states, and in making social and moral evaluations. Because counterfactual thinking underpins many cognitive capacities, it's important to determine when it develops. So far, results have been mixed, from children succeeding as young as two to as old as twelve. We present a new paradigm for studying counterfactual thinking: one that doesn't require counterfactual language, and teases apart hypothetical thinking (future-oriented) and counterfactual thinking (past-oriented). Using this paradigm, we find that 5-year-old children pass the test.

Keywords: counterfactual reasoning; hypothetical reasoning; development; social cognition.

Corresponding author: David Rose (davdrose@stanford.edu). All the data, study materials, preregistrations, and analysis code are available here: `https://github.com/cicl-stanford/counterfactual_development`

Introduction

David Lewis charged Hume with defining causation twice over: first, in terms of the regular succession of events, and second in terms of counterfactuals—had the first event not occurred, then neither would the second (Lewis, 1973). What Hume identified, and Lewis clarified, is the central role that counterfactuals play in determining what causes what. But counterfactuals aren't only important for thinking about causation (German, 1999; P. L. Harris, German, & Mills, 1996; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018). Counterfactual thinking is central in regret (e.g., Byrne, 2016; Coricelli & Rustichini, 2010), pretense (P. L. Harris, 1992; Leslie, 1987; Nichols & Stich, 2003), rule understanding (P. L. Harris, 2000; P. L. Harris & Núntez, 1996), mental state understanding (German & Nichols, 2003; Peterson & Bowler, 2000; Peterson & Riggs, 1999; Riggs & Peterson, 2014; Riggs, Peterson, Robinson, & Mitchell, 1998), and social and moral evaluation (Gautam & McAuliffe, 2024; Gautam, Owen Hall, Suddendorf, & Redshaw, 2023; Jaroslawska, McCormack, Burns, & Caruso, 2020; Kushnir, 2022; Pesowski, Denison, & Friedman, 2016; Wong, Cordes, Harris, & Chernyak, 2023; Zhao, Zhao, Gweon, & Kushnir, 2021).

Given the fundamental role that counterfactual thinking plays in cognition, many might find it surprising that we still don't understand *when* it emerges in development: estimates range from as early as age 2 (P. Harris, 1997) to as late as 12 (Rafetseder, Schwitalla, & Perner, 2013), and virtually everywhere in between (German & Nichols, 2003; Kominsky et al., 2021; McCormack, Ho, Gribben, O'Connor, & Hoerl, 2018; Nyhout & Ganea, 2019; Nyhout, Henke, & Ganea, 2017; Rafetseder, Cristi-Vargas, & Perner, 2010; Riggs et al., 1998). Despite the variety of methods used in these studies, one thing that many have in common is that they probe counterfactual thought by asking children counterfactual questions (e.g., Nyhout & Ganea, 2019, 2020; Nyhout et al., 2017; Rafetseder et al., 2010; Riggs et al., 1998). For example, after learning that Peter was in bed but got called to the Post Office to help put out a fire, children were asked "Where would Peter have been, had there not been a fire?" (e.g., Riggs et al., 1998). Three- and four-year-old children struggled to offer the correct response. Namely, that Peter would have still been in bed. Because counterfactual questions are linguistically complex, children may fail to understand them, and this would mask their ability to think counterfactually. We need a way to assess counterfactual thinking without counterfactual language. Additionally, we need a task that assesses genuine counterfactual thinking (Pearl, 2000; see Figure 1).

Counterfactual thinking is directed toward the past. It involves taking into account what actually happened, mentally traveling back in time to imagine a change to an event, and then simulating how this alternative would have played out. This kind of thinking is critical for judging causation and responsibility (Gerstenberg, 2024; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Kahneman & Tversky, 1982; Lagnado, Gerstenberg, & Zultan, 2013), and for giving explanations (Harding, Gerstenberg, & Icard, 2025; Hilton, 1990). *Hypothetical thinking*, in contrast, is directed toward the future. It involves simulating the consequences of taking (hypothetical) actions. This kind of thinking is critical for planning and decision-making (Gerstenberg, 2022; Sloman & Hagmayer, 2006). Finally, *conditional thinking* merely involves applying one's knowledge to reason from cause to effect, or from effect to cause (Skovgaard-Olsen, Stephan, & Waldmann, 2021).

Earlier work that argued for counterfactual thinking in 2-year-olds (P. Harris, 1997), was later interpreted to have shown that children succeeded by mere conditional thinking (Beck & Guthrie, 2011; Rafetseder et al., 2010). For example, to correctly answer that the leaves would not have fallen from the tree if there had been no wind, just requires thinking from cause to effect (see Figure 1). Similarly, a child may succeed in answering a counterfactual question through mere hypothetical thinking. Prior work argued that 4-year-olds can think counterfactually based on their performance in a blicket detector task (Nyhout & Ganea, 2019). Children first saw four blocks put on the detector one by one. Some turned the detector on, and some didn't. Later, the children saw pairs of blocks placed on the detector, and they were asked whether the detector would still be on without one of the blocks on it. Even though children were asked a counterfactual question, they could have succeeded by merely imagining what would happen if one of the blocks were put on the detector. In fact, they don't even need to imagine that as they've already seen each block being placed on the detector individually. So they can answer the counterfactual question correctly by remembering what happened.

A stronger test for counterfactual thinking involves simulating something one hasn't seen before. Moreover, to tease apart counterfactual from hypothetical thinking, one must learn something new from what actually happened that one couldn't have perfectly predicted before (Gerstenberg, 2022). Consider a situation where two causes, A and B, contribute to an outcome C. The status of B is initially uncertain and only gets revealed after A is observed. Here, the hypothetical question of *what would happen* to C without A (before knowing the status of B) yields a different answer than the counterfactual question of *what would have happened* to C without A (after knowing the status of B). We exploit this subtle but important distinction between future-oriented hypothetical thinking and past-oriented counterfactual thinking in our experimental paradigm.

Experiment overview

When do children begin to engage in counterfactual thinking? To answer this question, we need a task that distinguishes counterfactual thinking from hypothetical and conditional thinking (see Figure 1). And, to test the possibility that even young children are capable of counterfactual thinking, we need a task that doesn't probe counterfactual thinking with counterfactual language.

Figure 2 gives an overview of our experimental paradigm. In Experiment 1, Granny drops two different objects, such as an egg and an apple. Andy catches the egg and Suzy catches the apple. Not catching one (the egg) would have been worse than not catching the other (the apple). Participants are asked whether Andy or Suzy should get a thank-you sticker for catching the object (Granny only has one sticker). In Experiment 2, Granny drops two objects of the same kind. Again, both objects are caught and not catching one would have been worse (the egg above the ground) than not catching the other (the egg above the pillow). In both of these experiments, participants have visual access to the full scene. This means that they could in principle already compute what would happen if the person dropped the objects and then compare the outcome of this hypothetical simulation with what actually happened. To rule out this strategy, Experiment 3 initially shows a partial view of the scene, and then reveals the full scene only after the objects have been

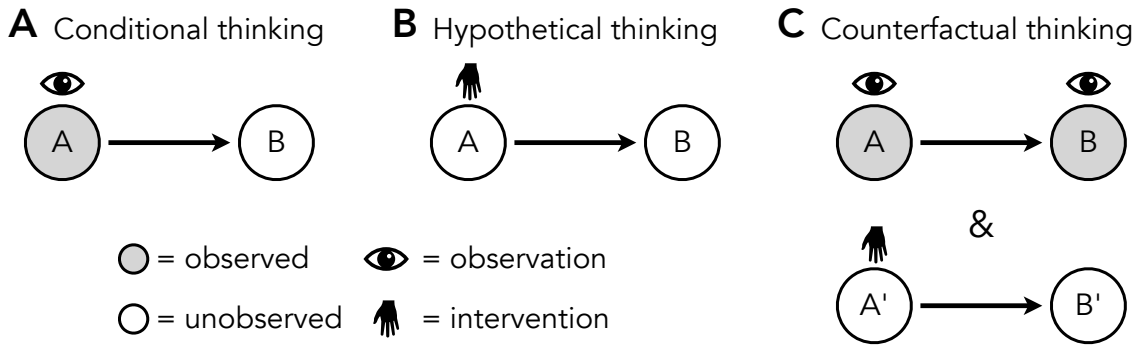

Figure 1

***Overview of Conditional, Hypothetical and Counterfactual Thinking**: **A** Conditional thinking involves drawing inferences about unobserved events from observed events. Note that one can reason from (observed) cause to (unobserved) effect (e.g., predicting where a falling ice cream cone will land in the sand), or from effect to cause (e.g., inferring from seeing the ice cream in the sand where it must have come from). **B** Hypothetical thinking involves considering the consequences of actions (e.g., predicting what would happen if the ice cream were dropped). **C** Counterfactual thinking involves considering how things could have played out differently (e.g. considering what would have happened if the ice cream had been dropped but not caught, when in fact it was caught). While both hypothetical thinking and counterfactual thinking involve considering the consequences of interventions, the key difference is whether it was observed what actually happened. Hypothetical thinking considers interventions in the future, whereas counterfactual thinking considers interventions in the past.*

caught. Here, counterfactual thinking is necessary to evaluate which outcome would have been worse as one cannot anticipate the possible outcomes in advance.

Experiment 1: Granny drops two different objects

Experiment 1 features scenarios where Granny drops two different objects (see Figure 4). We had two pre-registered hypotheses for children. The first was that children would be more likely to say that the individual should get the thank-you sticker who prevented the worse potential outcome from occurring. The second hypothesis was that the likelihood of choosing the person who prevented the worse potential outcome from occurring would increase with age. We also tested adults on the same materials. We hypothesized that they would be more likely to choose the individual who prevented the worse potential outcome from occurring.

Methods

All materials, data, pre-registrations, and analyses are available here: `https://github.com/cicl-stanford/counterfactual_development`

Figure 2

Experiment Overview*: In all experiments, Granny drops two objects, and each of them is caught. She only has one thank-you sticker. Who should she give it to? In Experiment 1, Granny drops two different objects. It would have been worse if the egg hadn't been caught than if the apple hadn't been caught. In Experiment 2, Granny drops two objects of the same kind (e.g., two eggs). Here, one of the objects would have landed on a soft pillow and the other on the floor. In Experiment 3, the full scene is revealed only after two objects of the same kind were caught. This means that participants cannot anticipate at the very beginning what would happen if Granny dropped the objects.*

Participants

Our final sample included 120 children, who were between the ages of 3 and 6, through Lookit who met our pre-registered inclusion criteria (*gender*: 60 female, 59 male, 1 no response/other; *language*: 114 English, 6 no response/other). 82 children were excluded for failing the warmup trial (see below).[1] Each age group included 30 participants. Families

[1]To be included, participants need to pass the warmup trial. Since our task involved quite a number of friends' names and objects across trials, we wanted to ensure they could track this. To accommodate younger participants, we allowed children to respond by pointing, with parents verbally reporting their child's answer.

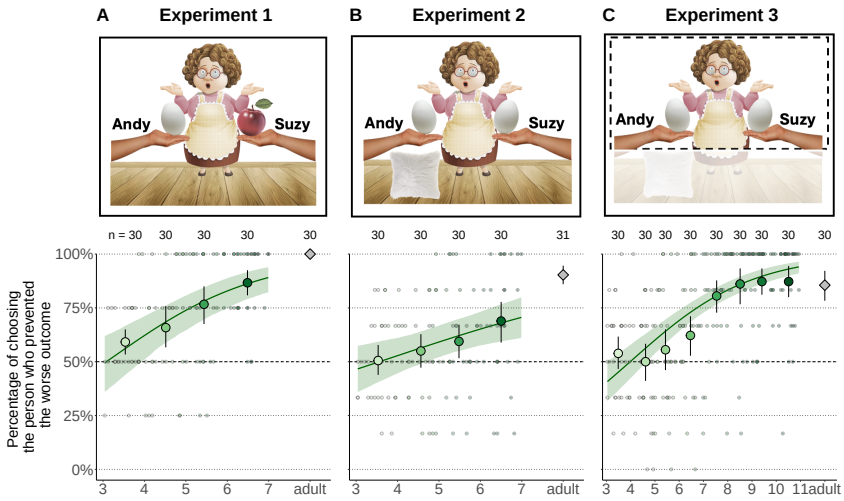

Figure 3

***Overall Experimental Results**: Percentage of participants who chose the person who prevented the worse outcome from happening across the three experiments. Large points show means for each age group with 95% bootstrapped confidence intervals. Small points show individual responses averaged across trials. Dashed lines indicate chance performance. The number of participants in each age group is shown at the top. **A** In Experiment 1, when two different objects are dropped and prevented from falling onto the ground, children between the ages of 3 and 4 are already inclined to think the person who prevented the worse outcome from occurring (e.g., catching the egg) should receive the thank-you sticker. **B** In Experiment 2, when two of the same kinds of objects are dropped and prevented from falling onto the ground, children between the ages of 5 and 6 are inclined to think the person who prevented the worse outcome from occurring (e.g., catching the egg with the floor underneath) should receive the thank-you sticker. **C** In Experiment 3, where only after the objects are caught is it revealed to participants where the objects would have landed, children between the ages of 6 and 7 are inclined to think the person who prevented the worse outcome from occurring (e.g., catching the egg with the floor underneath) should receive the thank-you sticker.*

received $5. Additionally, we recruited 30 adult participants through Prolific who met our pre-registered inclusion criteria (*age*: M = 35, SD = 13; *gender*: 21 female, 9 male; *race*: 21 White, 5 no response/other, 3 Black, 1 Asian). Participants were compensated at a rate of $12 per hour.

Materials

We created six trials, where four of these were test trials, one was a warm-up and one was a post-test trial. The warm-up trial showed Granny, a plant and a lamp on opposite

In these cases, we ensured parents remained neutral and did not influence their child's responses. Despite these accommodations, many 3-year-olds still could not meet this basic comprehension criterion and were therefore excluded from the main analysis.

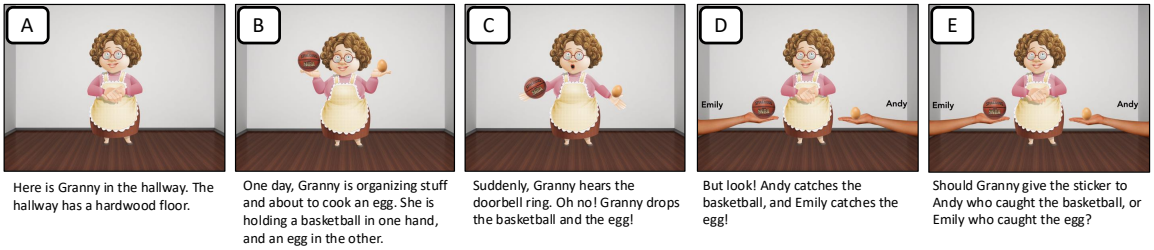

Figure 4

Experiment 1 example trial*: Each panel shows what participants saw and what they were told.* ***A*** *Participants first see that Granny is in a hallway. And they are told that the hallway has hardwood floors.* ***B*** *Granny is holding a basketball and egg.* ***C*** *Granny drops the objects after being startled by the doorbell.* ***D*** *The two objects are caught by Emily and Andy.* ***E*** *Participants are asked who Granny should thank.*

sides of her, and two hands, with one by the plant and one by the lamp. The materials for the test trials depicted Granny holding two different objects in different situations. These are shown in Figure 5. The post-test trial showed Granny outside by two piles of leaves, where one pile was smaller than the other.

Procedure

The experiment for children was programmed using Lookit (Scott & Schulz, 2017) and for adults it was programmed using jsPsych (De Leeuw, 2015). Both children and adults were tested asynchronously.

Participants began with the warm-up trial. They were first introduced to Granny. They were then told that Granny had a plant and a lamp. Each of these was on opposite sides of Granny and both wiggled when they were introduced. Then participants were told that they would meet two of Granny's friends. They were shown an arm on the plant side of the screen, told that this is Benji and that Benji is by the plant. And they were shown an arm on the lamp side of the screen, told that this is Harry and that Harry is by the lamp. Participants were then asked who is by the plant and who is by the lamp. Children responded out loud. Adults clicked one of two buttons that were labeled "Benji" and "Harry".

Participants then proceeded to the four test trials, where each involved Granny in different situations, holding different objects and dropping them. Both objects are then caught. An example of the procedure for the basketball and egg trial is shown in Figure 4. In all test trials, after the two objects were dropped and caught, participants were told that Granny only has one thank you sticker. Participants were asked who Granny should give it to. Children responded out loud and adults selected one of two buttons with the name of the individuals who caught the objects.

Lastly, participants completed a post-test trial. They were told that Granny has two of her friends help her to rake leaves. They are then shown two piles of leaves, one of which was much smaller than the other, and asked who Granny should thank.

Design

The order of the test trials was randomized within participants. In addition, for all test trials, the side of the screen that the dropped objects were on and names of the individuals who caught the objects was randomized across participants. The position of the object that would result in a worse outcome was counterbalanced across the four test trials—with two being on the left side and two being on the right side.

Results

Overall results are shown in Figure 3, and broken down by trial in Figure 5. As can be seen, the pattern is largely consistent across individual trials.

For all three experiments reported here, we coded responses so that 1 = selecting the person who prevented the worse outcome and 0 = selecting the other person. We then fit separate Bayesian regression models for children and adults with random intercepts for both participants and trials.[2]

As predicted, we found that both children and adults were more likely to select the individual who prevented the worse outcome from happening as the person who should receive the thank-you sticker. In a model that only featured the global intercept as a fixed effect and was fitted to the children's responses, the posterior estimate for the intercept was positive, and the 95% credible interval excluded 0, $\beta = 0.73$, 95% credible interval (CrI) =

[2] All Bayesian models were written in `Stan` (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019).

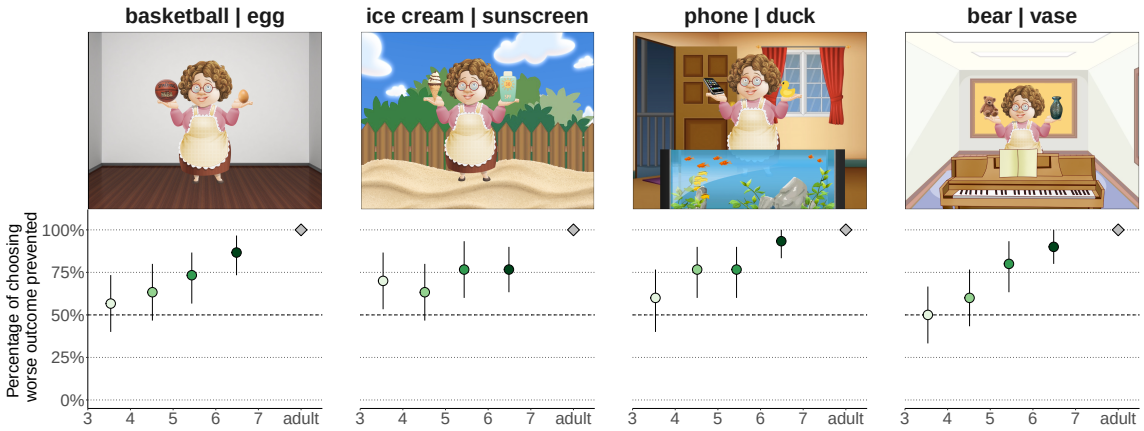

Figure 5

Experiment 1*: Percentage of participants who chose the person who prevented the worse outcome from happening in each of the four test trials. Points show means for each age group with 95% bootstrapped confidence intervals. Dashed lines indicate chance performance. Even 3-year-olds, the youngest children we tested, performed above chance overall. The pattern of performance was similar across the four different scenarios.*

$[0.65, 0.80]$. Adults chose the individual who prevented the worse outcome from occurring 100% of the time.

As predicted, we also found that older children were more likely to select the individual who prevented the worse outcome from happening. In a model that featured a global intercept and age as fixed effects, there was a credible effect of age $\beta = 0.55$, 95% CrI = $[0.34, 0.75]$.[3]. We also explored at what age children perform above chance. We found that the lower bound of the 95% credible interval from the regression model exceeds the chance level at 3.95 years.

Discussion

When two different objects are dropped, children are more inclined to select the person who prevented the worse outcome from occurring as the person who should receive a thank-you sticker. For instance, if Granny drops an egg and a basketball and both are caught before they land on the floor, children are more likely to think the thank-you sticker should go to the person who caught the egg. This suggests that they appreciate that the person who caught the egg should receive the sticker, because had they not caught it, it would have led to a worse outcome than not catching a basketball. In other words, children may be engaging in counterfactual thinking to decide who to thank. If so, then this suggests a procedure to assess the capacity for counterfactual thinking without relying on whether children understand counterfactual language.

It may be, however, that our task doesn't yet provide a clear way to assess the capacity for counterfactual thinking. There are other explanations of our findings available that don't appeal to counterfactual thinking. For example, it could be that children are deciding who to select by either thinking about what object they like or what object Granny likes and then deciding that the person who caught the liked object should receive the thank-you sticker. In other words, children, in deciding who should receive the thank-you sticker, may not even consider which outcome would have been worse. They might move straight from a judgment about what they like to who Granny should thank. Or they might move straight from a judgment about what Granny likes to who should receive the thank-you sticker. To address this, we examine, in our next experiment, whether 3–6 year-old children continue to select the person who prevented the worse outcome from occurring when two of the *same objects* are dropped but where one's not being caught would have resulted in a worse outcome.

Experiment 2: Granny drops the same objects

In Experiment 2, Granny drops two objects of the same kind. This means that children cannot give the correct answer anymore by considering merely which object they themselves or Granny might like better. The pre-registered hypotheses were the same as in Experiment 1.

[3]We adopt the convention of calling an effect "credible" when the 95% credible interval of the posterior distribution for the parameter of interest excludes 0.

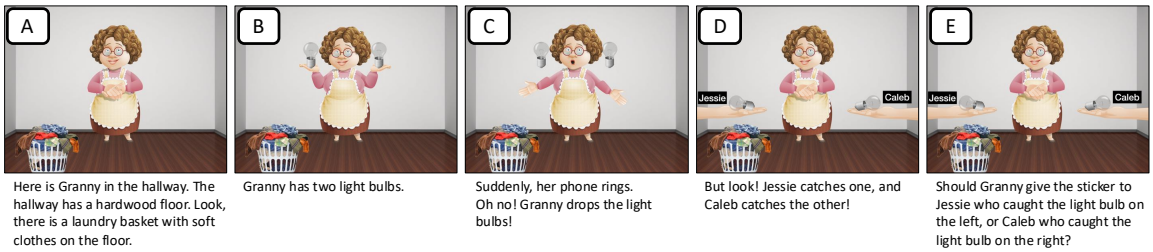

Figure 6

***Experiment 2 example trial**: Each panel shows what participants saw and what they were told. **A** Participants first see that Granny is in a hallway. And they are told that the hallway has hardwood floors, and that a laundry basket is on the floor. **B** Granny is holding two lightbulbs. **C** Granny drops the objects after being startled by the phone ringing. **D** The two objects are caught by Jessie and Caleb. **E** Participants are asked who Granny should thank.*

Methods

Participants

Our final sample included 120 children, who were between the ages of 3 and 6, through Lookit who met our pre-registered inclusion criteria (*gender*: 65 female, 55 male; *language*: 118 English, 2 no response/other). 74 participants were excluded for failing to pass the warmup trial. Each age group included 30 participants. Families received $5. Additionally, we recruited 31 adult participants through Prolific who met our pre-registered inclusion criteria (*age*: M = 39, SD = 14; *gender*: 19 female, 11 male, 1 no response/other;

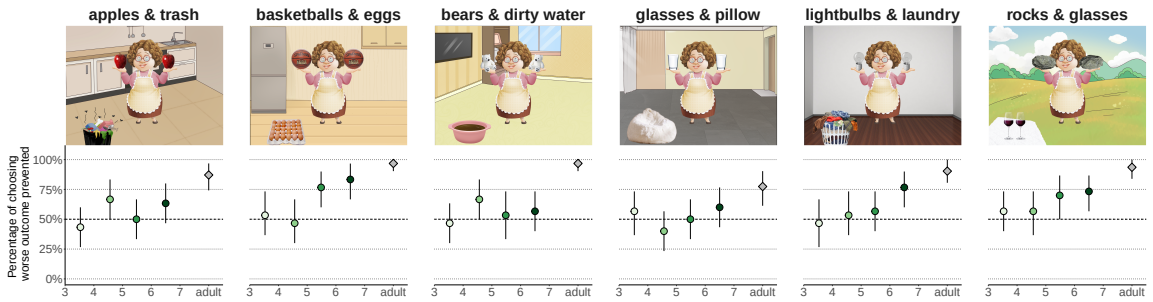

Figure 7

***Experiment 2**: Percentage of participants who chose the person who prevented the worse outcome from happening in each of the six test trials. Points show means for each age group with 95% bootstrapped confidence intervals. Dashed lines indicate chance performance. Children aged 5 years and older performed above chance overall. There was some variability in performance across the different trials. Notably, even adult participants had mixed intuitions about the glasses & pillow trial.*

race: 20 White, 7 Black, 2 no response/other, 1 Asian). Participants were compensated at a rate of $12 per hour.

Materials

The warm-up trial was the same as in Experiment 1. The materials for the test trials, however, differed from Experiment 1 in that they depicted Granny holding two of the same objects in different situations. All six test trials are shown in Figure 7.

Procedure

The procedure was the same as in Experiment 1. An example trial is shown in Figure 6.

Design

The design was the same as in Experiment 1.

Results

Overall results are shown in Figure 3, and broken down by trial in Figure 7. There was some variability in performance across the trials. For example, even adult participants were less certain about the glasses & pillow trial. One possible reason is that while the pillow might prevent the glass from breaking, it would arguably be worse if the glass broke on the pillow rather than on the ground.

As predicted, we found that both children and adults were more likely to select the individual who prevented the worse outcome from happening as the person who should receive the thank-you sticker (children: $\beta = 0.37,\ 95\%\ \text{CrI} = [0.07, 0.72]$, adults: $\beta = 2.44,\ 95\%\ \text{CrI} = [1.49, 3.44]$).

As predicted, we also found that older children were more likely to select the individual who prevented the worse outcome from happening, $\beta = 0.26,\ 95\%\ \text{CrI} = [0.09, 0.44]$. Children performed above chance at 4.71 years.

Discussion

Even when two of the same objects are dropped, children are more inclined to select the person who prevented a worse outcome from occurring as the person who should receive the thank-you sticker. For instance, if two glasses are dropped, and one would have fallen onto a hard floor if not caught and the other would have fallen onto a soft pillow, children are more inclined to think that the person who prevented the glass from falling on the floor should receive the thank-you sticker.

When two different objects were dropped, three-year-old children were already selecting the person who prevented the worse outcome at rates above chance; when two of the same objects were dropped, it wasn't until around five years of age that children began selecting the person who prevented the worse outcome from occurring at rates above chance. This suggests that when two different objects were dropped (Experiment 1), children's judgments about what they like or what Granny likes may have influenced their responses. But here, since the objects are the same, it is more challenging to rely on this when responding. Instead, children need to simulate what would have happened had the objects not been

caught, determine which outcome would have been worse, and then decide to select the person who prevented the worse outcome from occurring as the one who should receive the thank-you sticker. There is, however, a remaining issue.

In Experiment 2, participants were able to see the whole scene from the very beginning. This means that they were able to simulate what would happen if Granny dropped one object or the other (before any of this happened). For example, when Granny holds a glass over a pillow and another glass over the floor, they can simulate that dropping the glass over the floor would be worse than dropping the glass over the pillow. They can then compare the outcome of these hypothetical simulations to what actually happened, and make the judgment that catching the glass over the floor is more deserving of a reward.

To address this concern, Experiment 3 reveals the full scene only *after* the objects were caught. This subtle difference makes it such that hypothetical thinking doesn't suffice anymore because one cannot anticipate what would happen. As such, this experiment provides a stronger test for children's ability to think counterfactually.

Experiment 3: Granny drops the same objects and the full scene is only revealed later

In Experiment 3, Granny again drops two objects of the same kind, but this time the full scene is only revealed after both objects were caught. Our pre-registered hypotheses were the same as in Experiments 1 and 2.

Methods

Participants

We expanded the age range from previous experiments since we expected the task to be more challenging. Our final sample included 240 children aged 3 to 10 through Lookit who met our pre-registered inclusion criteria (*gender*: 137 female, 102 male, 1 no response/other; *language*: 235 English, 5 no response/other). 101 participants were excluded for failing to pass the warmup trial. Each age group included 30 participants. Families received $5. Additionally, we recruited 30 adult participants through Prolific who met our pre-registered inclusion criteria (*age*: M = 39, SD = 12; *gender*: 20 female, 9 male, 1 no response/other; *race*: 23 White, 5 Black, 1 no response/other, 1 Asian).

Materials

The warm-up trial was the same as in our previous experiments. The materials were the same as in Experiment 2.

Procedure

The procedure was the same as in our previous experiments except that after the objects were caught, it was revealed where they would have fallen. An example trial is shown in Figure 8.

Design

The design was the same as in Experiment 2.

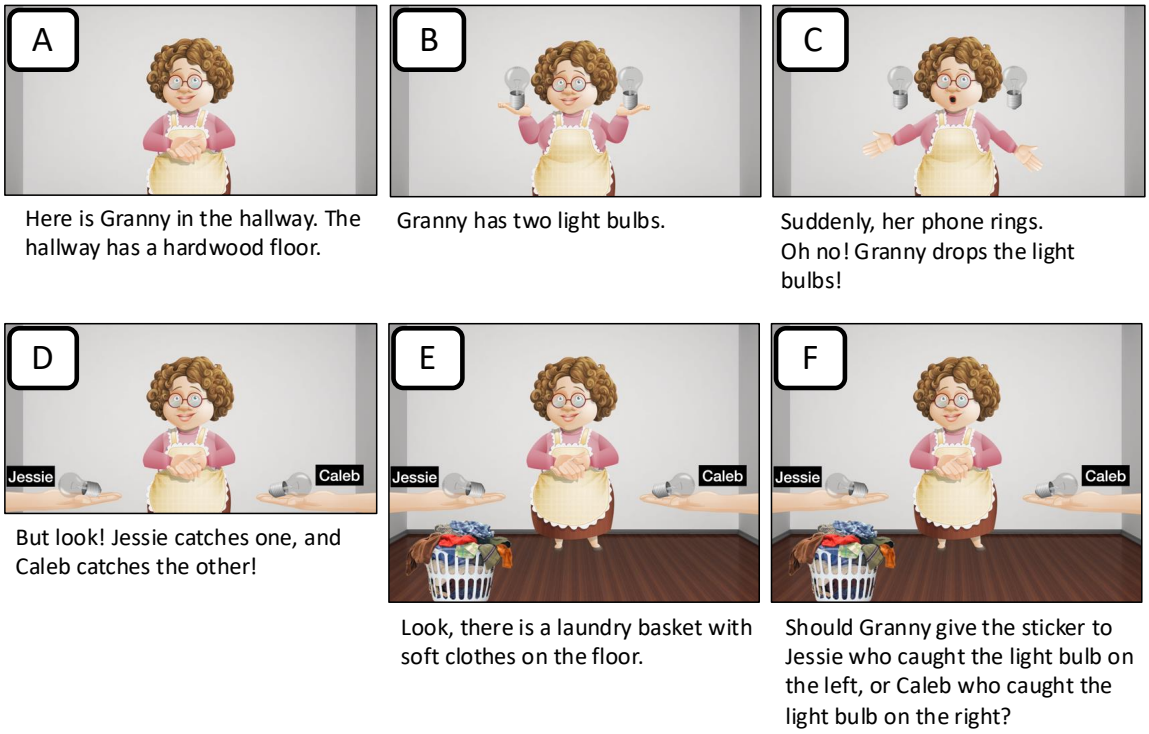

Figure 8

***Experiment 3 example trial**: Each panel shows what participants saw and what they were told. Notice that unlike in Experiment 2, participants didn't view the full scene at the beginning. The fact that there was a laundry basket underneath one of the light bulbs was only revealed after the light bulbs were caught. This means that participants could not anticipate what would happen if the light bulbs were dropped at the beginning of the scene, but instead had to consider what would have happened if the light bulbs hadn't been caught at the end of the scene.*

Results

Overall results are shown in Figure 3, and broken down by trial in Figure 9. The developmental pattern was largely consistent across the trials.

As predicted, we found that both children and adults were more likely to select the individual who prevented the worse outcome from happening as the person who should receive the thank-you sticker (children: $\beta = 1.19$, 95% CrI $= [0.95, 1.45]$, adults: $\beta = 2.78$, 95% CrI $= [0.74, 4.99]$).

As predicted, we also found that older children were more likely to select the individual who prevented the worse outcome from happening, $\beta = 0.40$, 95% CrI $= [0.31, 0.50]$. Children performed above chance at 5.02 years.[4]

[4]For this analysis, we restricted the age range from 3- to 6-year-old children, as in Experiments 1 and 2. Using the full age range, we find that children perform above chance at 4.67 years old. This shift in the slope of the regression line is due to the high performance of older children. We would expect a similar shift

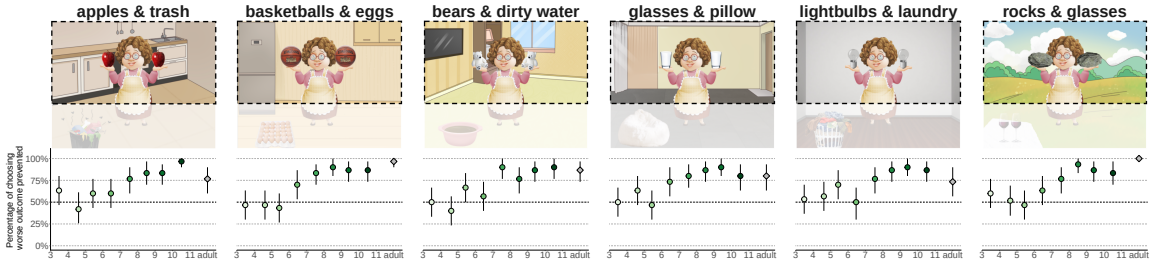

Figure 9

Experiment 3*: Percentage of participants who chose the person who prevented the worse outcome from happening in each of the six test trials. Points show means for each age group with 95% bootstrapped confidence intervals. Dashed lines indicate chance performance. Children aged 6 years and older performed above chance overall. The pattern of performance was largely similar across the trials, with a notable increase in performance around age 7.*

Discussion

Even when two of the same objects are dropped, caught, and it is later revealed where the objects would have fallen, children are more inclined to select the person who prevented a worse outcome from occurring as the one who should receive the thank-you sticker. When children already knew where two objects would land before seeing them dropped (Experiment 2), it wasn't until around five years of age that they began selecting at above chance rates the person who prevented the worse outcome from occurring as the one to thank. However, their responses might have resulted from hypothetical, not counterfactual, thinking. When the full scene is visually available, children could simulate a hypothetical situation where the objects are dropped and determine which would result in a worse outcome. Comparing these two hypothetical outcomes to what actually happened is sufficient for deciding who should get the thank-you sticker. In Experiment 3, children cannot rely on this strategy as the full scene is only revealed after the objects were caught. Now, children need to simulate the counterfactual situation of what would have happened if the objects hadn't been caught. In this version of the task, we find that around 5 years old, children reliably select the person who prevented the worse outcome from occurring as the one who should be thanked.

General Discussion

Counterfactual thinking is a fundamental cognitive capacity. Prior work on how and when children develop this ability has yielded mixed results. This is likely due to the different ways in which counterfactual thinking has been assessed. We argue that some of the prior work showing that children succeed early may be better explained by assuming that they engaged in hypothetical thinking, but not counterfactual thinking. We developed a paradigm that allows us to tease apart hypothetical and counterfactual thinking. Importantly, this paradigm doesn't require children to understand counterfactual language.

---

for Experiments 1 and 2, if the same age range was tested as in Experiment 3.

Because counterfactual language is complicated, paradigms that use such language could mask an early underlying ability to think counterfactually.

Our paradigm relies on children's intuitive understanding of the physical and social world. To make sure that children have the relevant causal understanding in place, we focused on simple physical events: dropping and catching things. Catching an object means that one prevented it from falling to the ground, and recognizing that an outcome was prevented requires counterfactual thinking. Even young children have plenty of experience with things dropping, and with adults expressing their emotions about that. How bad it is to drop something varies: dropping a phone into an aquarium is worse than dropping it onto a cushion.

We asked participants to predict who should get a thank-you sticker when two characters each caught an object. Saying "thank you" is a social act that children learn early on. Importantly, this prompt gets at counterfactual thinking without requiring counterfactual language. We don't need to explicitly ask children what would have happened if an object hadn't been caught. But predicting who should be thanked more requires simulating who prevented the worse outcome from happening. Moreover, while in much prior work on counterfactual thinking, children can answer counterfactual questions by recalling an event they had seen earlier, in our setting children never see an object hitting the ground. They need to mentally simulate this counterfactual outcome.

Using this paradigm, we examined the development of hypothetical and counterfactual thinking across three experiments. In Experiment 1, when Granny drops two different objects, such as an ice cream cone and bottle of sunscreen, children as young as three say that Granny should thank the person who prevented the worse outcome from occurring—the person who caught the ice cream cone. Because Granny drops two different objects, it's possible that participants just pick the character who catches the object they think Granny likes better.

In Experiment 2, we rule out this strategy by making Granny drop the same two objects. For example, when Granny drops two glasses and one would have fallen on the floor, and the other on a pillow, around 5 years old, children are inclined to select the person who prevented the worse outcome from occurring—the person who caught the glass that would have fallen onto the floor—as the person who should receive the thank-you sticker. However, because participants in these experiments can see the full scene from the very beginning, the results are consistent with children engaging merely in hypothetical thinking—simulating from the very beginning that dropping one glass would be worse than the other, and comparing that to what actually happened.

Experiment 3 provides a more stringent test for counterfactual thinking because the full scene is revealed only after the two objects are caught. This way, participants cannot anticipate in advance what would happen. When Granny drops two glasses and it is revealed after they were caught that one would have fallen on the floor, and the other onto a pillow, around 5 years old, children think that the person who prevented the glass from falling onto the floor should get the thank-you sticker.

Our paradigm teases apart counterfactual thinking from hypothetical thinking and we show that when hypothetical thinking is sufficient, children succeed at an earlier age than when counterfactual thinking is necessary. Even though our task doesn't require counterfactual language, it does still require language. Future work could test implicit counterfactual

thinking without language, for example by using eye-tracking (see Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017). It is possible that such measures may find success at an earlier age than we did here. Pinning down the age at which children succeed in counterfactual thinking is important because it helps us better understand how different cognitive capacities are related to one another. If it's indeed the case that children only learn to reason counterfactually around five years of age, then this means that there are a lot of cognitive, physical, and social capacities for which counterfactual thinking isn't necessary. More work is needed to better understand what role the capacity for counterfactual thinking plays in how children learn and develop. For example, how does counterfactual thinking matter for learning, pragmatic language understanding, planning, and decision-making?

Conclusion

Counterfactual thinking is a hallmark of human intelligence. Much prior work has investigated how this capacity develops and found mixed results, from children succeeding as young as two to as old as twelve. We presented a new paradigm for studying counterfactual thinking: one that doesn't require counterfactual language, and teases apart hypothetical and counterfactual thinking. Using this paradigm, we find that 5-year-old children pass the test.

Acknowledgment

We thank Matan Mazor for suggesting the idea of occluding part of the scene in Experiment 3 in order to tease apart counterfactual and hypothetical thinking. TG was supported by grants from Stanford's Human-centered Artificial Intelligence Institute (HAI) and from Cooperative AI.

References

Beck, S. R., & Guthrie, C. (2011). Almost thinking counterfactually: Children's understanding of close counterfactuals. *Child development*, *82*(4), 1189–1198.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*, 135–157.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Coricelli, G., & Rustichini, A. (2010). Counterfactual thinking and emotions: regret and envy learning. *Philosophical Transactions of the Royal society B: Biological sciences*, *365*(1538), 241–247.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*, 1–12.

Gautam, S., & McAuliffe, K. (2024). Why imagining what could have happened matters for children's social cognition. *WIREs: Cognitive Science*, *15*(1).

Gautam, S., Owen Hall, R., Suddendorf, T., & Redshaw, J. (2023). Counterfactual choices and moral judgments in children. *Child Development*, *94*(5), e296–e307.

German, T. P. (1999). Children's causal reasoning: Counterfactual thinking occurs for 'negative'outcomes only. *Developmental Science*, *2*(4), 442–457.

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

Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, *377*(1866), 20210339.

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, *28*(10), 924–936.

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*(6), 936–975.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from `https://doi.org/10.1177%2F0956797617713053` doi: 10.1177/0956797617713053

Harding, J., Gerstenberg, T., & Icard, T. (2025). A Communication-First Account of Explanation. *arXiv*. Retrieved from `https://arxiv.org/abs/2505.03732`

Harris, P. (1997). On realizing what might have happened instead. *Polish Quarterly of Developmental Psychology*, *3*, 161–176.

Harris, P. L. (1992). From simulation to folk psychology: the case for development. *Mind & Language*.

Harris, P. L. (2000). *The work of the imagination.* Blackwell Publishing.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233–259.

Harris, P. L., & Núntez, M. (1996). Understanding of permission rules by preschool children.

Child development, *67*(4), 1572–1591.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81.

Jaroslawska, A. J., McCormack, T., Burns, P., & Caruso, E. M. (2020). Outcomes versus intentions in fairness-related decision making: School-aged children's decisions are just like those of adults. *Journal of experimental child psychology*, *189*, 104704.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kominsky, J. F., Gerstenberg, T., Pelz, M., Sheskin, M., Singmann, H., Schulz, L., & Keil, F. C. (2021). The trajectory of counterfactual simulation in development. *Developmental Psychology*, *57*(2), 253.

Koskuba, K., Gerstenberg, T., Gordon, H., Lagnado, D. A., & Schlottmann, A. (2018). What's fair? how children assign reward to members of teams with differing causal structures. *Cognition*, *177*, 234-248.

Kushnir, T. (2022). Imagination and social cognition in childhood. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(4), e1603.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind.". *Psychological review*, *94*(4), 412.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, *45*, 1–9. Retrieved from `https://doi.org/10.1016%2Fj.cogdev.2017.10.001` doi: 10.1016/j.cogdev.2017.10.001

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4- and 5-year-olds. *Cognition*, *183*, 57-66. doi: 10.1016/j.cognition.2018.10.027

Nyhout, A., & Ganea, P. A. (2020). What is and what never should have been: Children's causal and counterfactual judgments about the same events. *Journal of Experimental Child Psychology*, 104773. Retrieved from `http://dx.doi.org/10.1016/j.jecp.2019.104773` doi: 10.1016/j.jecp.2019.104773

Nyhout, A., Henke, L., & Ganea, P. A. (2017). Children's counterfactual reasoning about causally overdetermined events. *Child Development*. doi: 10.1111/cdev.12913

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

Pesowski, M. L., Denison, S., & Friedman, O. (2016). Young children infer preferences from a single action, but not if it is constrained. *Cognition*, *155*, 168–175.

Peterson, D. M., & Bowler, D. M. (2000). Counterfactual reasoning and false belief understanding in children with autism. *Autism*, *4*(4), 391–405.

Peterson, D. M., & Riggs, K. J. (1999). Adaptive modelling and mindreading. *Mind & Language*, *14*(1), 80–112.

R Core Team. (2019). R: A language and environment for statistical computing [Computer

software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child development*, *81*(1), 376–389.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404. doi: 10.1016/j.jecp.2012.10.010

Riggs, K. J., & Peterson, D. M. (2014). Counterfactual thinking in pre-school children: Mental state and causal inferences. In *Children's reasoning and the mind* (pp. 87–99). Psychology Press.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, *13*(1), 73–90.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

Skovgaard-Olsen, N., Stephan, S., & Waldmann, M. (2021). Conditionals and the hierarchy of causal queries. *Journal of Experimental Psychology: General*, *150*(12), 2472–2505.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, *10*(9), 407–412.

Wong, A., Cordes, S., Harris, P. L., & Chernyak, N. (2023). Being nice by choice: The effect of counterfactual reasoning on children's social evaluations. *Developmental Science*, *26*(6), e13394.

Zhao, X., Zhao, X., Gweon, H., & Kushnir, T. (2021). Leaving a choice for others: Children's evaluations of considerate, socially-mindful actions. *Child development*, *92*(4), 1238–1253.