Probing counterfactual thinking 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 fundamental cognitive capacity that underlies many aspects of our everyday lives; it allows us to learn from past mistakes, evaluate our own and others' actions, and imagine a world beyond the here and now. Yet, prior work has yielded a strikingly wide developmental window for the onset of counterfactual thinking: as early as 2, and as late as 12. There are at least two reasons for this: reliance on counterfactual language (which can underestimate children's competence), and a failure to distinguish counterfactual thinking from hypothetical thinking (which can overestimate children's competence). The current work presents a novel paradigm for probing genuine counterfactual thinking that does not require counterfactual language. After watching a scenario where Granny drops two items that are caught by two different characters, participants are asked which of the two characters Granny should thank. Across three experiments that implement different versions of the task to rule out alternative accounts, we find that the capacity for genuine counterfactual thinking may be present by around age 5, while younger children may succeed on tasks that can be solved via hypothetical thinking. By offering an intuitive and practical method for assessing counterfactual thinking without counterfactual language, the current work opens up a range of empirical questions about the interplay between the development of counterfactual thinking and other cognitive capacities.

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 counterfactual thinking plays in determining what causes what. Importantly, its role isn't just restricted to causation (German, 1999; P. L. Harris, German, & Mills, 1996; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018; Smallman & McCulloch, 2012; Summerville, 2011); prior work suggests that counterfactual thinking may also support our ability to feel regret and learn from past mistakes (Byrne, 2016; Camille et al., 2004; Coricelli & Rustichini, 2010; Roese, 1994; Smallman & McCulloch, 2012; Summerville, 2011), engage in pretense (P. L. Harris, 1992; Leslie, 1987; Nichols & Stich, 2003), and make social and moral judgments (Gautam & McAuliffe, 2024; Gautam, Owen Hall, Suddendorf, & Redshaw, 2023; Gerstenberg, 2024; Jaroslawska, McCormack, Burns, & Caruso, 2020; Koskuba et al., 2018; Kushnir, 2022; Lagnado, Gerstenberg, & Zultan, 2013; Pesowski, Denison, & Friedman, 2016; Stanley, Cabeza, Smallman, & De Brigard, 2021; Wong, Cordes, Harris, & Chernyak, 2023; Zhao, Zhao, Gweon, & Kushnir, 2021).

Despite the fundamental role that counterfactual thinking plays in cognition, we still don't fully understand *when* it emerges in development: prior work has estimated that it might be as early as age 2 (P. Harris, 1997) or as late as age 12 (Rafetseder, Schwitalla, & Perner, 2013), and virtually anywhere 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, Peterson, Robinson, & Mitchell, 1998). Why has prior research failed to find converging evidence? There are at least two possible reasons: first, many studies use complex counterfactual language to probe counterfactual thinking; second, many studies fail to rule out conditional or hypothetical thinking as an alternative strategy.

While individual studies vary in their methods, most rely on counterfactual language to probe counterfactual thinking (e.g., Nyhout & Ganea, 2019, 2020; Nyhout et al., 2017; Rafetseder et al., 2010; Riggs et al., 1998), though see Amsel and Smalley (2014); Beck and Crilly (2009); Guttentag and Ferrell (2004); Jones, Nelson, Gautam, and Redshaw (2025); McCormack, Feeney, and Beck (2020); McCormack, O'Connor, Beck, and Feeney (2016); O'Connor, McCormack, and Feeney (2012); Rafetseder and Perner (2012); Weisberg and Beck (2012). 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?" (Riggs et al., 1998). Understanding the question and offering the correct response—that Peter would have still been in bed—requires comprehending and producing linguistically complex counterfactual language. In English, for instance, counterfactuals are often expressed using conditional sentence structure (if-then) with past perfect tense (a verb form used to describe a past action that occurred before another past action) and modal verbs (e.g., "would"). Given that the cluster of linguistic abilities required to understand counterfactuals are relatively late emerging (Leahy & Carey, 2020; Ozturk & Papafragou, 2015; Shtulman & Phillips, 2018), prior work may have inadvertently masked children's competence, leading to age estimates that are relatively high.

Other work may have overestimated children's competence by conflating counterfac-

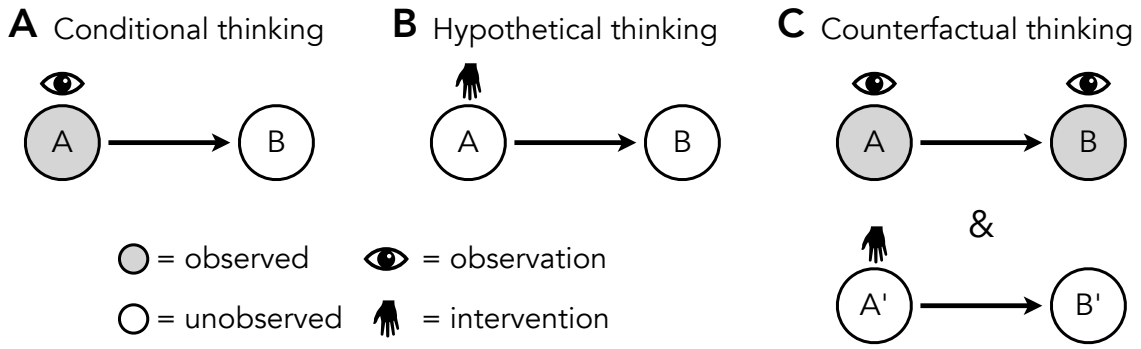

Figure 1

***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 basketball will land on the floor), or from effect to cause (e.g., inferring where a basketball on the floor must have come from). **B** Hypothetical thinking involves considering the consequences of actions (e.g., predicting what would happen if a basketball were dropped). **C** Counterfactual thinking involves considering how things could have played out differently (e.g., considering what would have happened if the basketball 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 what actually happened was observed. Hypothetical thinking considers interventions in the future, whereas counterfactual thinking considers interventions in the past.*

tual thinking with other types of reasoning. Figure 1 contrasts conditional, hypothetical, and counterfactual thinking. *Conditional thinking* usually involves reasoning from cause to effect, or from effect to cause (Skovgaard-Olsen, Stephan, & Waldmann, 2021). This kind of thinking underlies basic causal reasoning (Sloman & Lagnado, 2015). *Hypothetical thinking* is directed toward the *future*; it involves simulating the consequences of taking (hypothetical) actions. This kind of thinking underlies planning and decision making (Sloman & Hagmayer, 2006). *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 forward to infer how this alternative would have played out. Counterfactual thinking is more complex in that it combines elements of conditional thinking (taking into account what happened) with hypothetical thinking (simulating the consequences of intervening; see Gerstenberg, 2022; Pearl, 2000). This kind of thinking underlies, among other things, judging causation and attributing responsibility (Gerstenberg, 2024; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Hilton, 1990; Kahneman & Tversky, 1982; Lagnado et al., 2013).

Probing genuine counterfactual thinking is challenging. Even in research with adult participants, there are only a handful of counterfactual tasks where performance cannot be explained as reflecting mere hypothetical or conditional thinking (see Gerstenberg, 2022). Similar issues arise in developmental research. For instance, earlier work that argued for

counterfactual thinking in 2-year-olds (P. Harris, 1997) was later interpreted as success via mere conditional thinking (Beck & Guthrie, 2011; Rafetseder et al., 2010). Correctly answering that the leaves would not have fallen from the tree if there had been no wind, just requires conditional thinking from cause to effect (see Figure 1a).

Similarly, the results of a study that claimed to have found counterfactual thinking in 4-year-olds (Nyhout & Ganea, 2019) can be explained by hypothetical thinking. In one version of the task, children first saw three blocks placed on a "blicket detector" one by one, only some of which activated the detector. Then, after seeing two of those blocks sequentially placed on the detector (such that both blocks remain on the detector), 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. Since children saw each block being placed on the detector individually they could have also correctly answered the counterfactual question by simply remembering what happened, rather than counterfactually imagining what would have happened. Thus, by measuring conditional or hypothetical thinking instead of genuine counterfactual thinking, prior work may have overestimated children's competence.

In sum, the divergent findings in prior work can be attributed to at least two culprits: the use of counterfactual language (which may have raised the estimated age of success) and a failure to distinguish genuine counterfactual thinking from hypothetical or conditional thinking (which may have lowered the estimated age of success). In fact, many studies involve both: asking counterfactual questions about scenarios that can be resolved by engaging in hypothetical or conditional thinking (e.g., P. Harris, 1997; P. L. Harris et al., 1996; Nyhout & Ganea, 2019, 2020; Nyhout et al., 2017; Rafetseder et al., 2010; Riggs et al., 1998). In what follows, we introduce a novel paradigm that address both issues and use it to characterize the development of counterfactual thinking in young children.

Experiment overview

When do children begin to engage in counterfactual thinking? First, we need a task that doesn't rely on counterfactual language. To this end, we developed a task that involves watching a simple scenario where a character (Granny) drops two items—each of which is caught by two different characters—and answering a question about who should receive a thank-you sticker. This task does not require counterfactual language, in either comprehension or production; children simply need to understand that only one character can receive a sticker and choose one of the two. Second, we need a task that distinguishes counterfactual thinking from hypothetical and conditional thinking (see Figure 1). To this end, we designed different versions of the task that tap into different kinds of reasoning. Figure 2 gives an overview of our experimental paradigm and the differences between experiments.

In Experiment 1 (Figure 2a), Granny drops two different objects, such as a basketball and an egg. Critically, the anticipated consequence of one event is worse than the other (e.g., the basketball would bounce but the egg would break). After seeing that Andy catches the basketball and Suzy catches the egg, participants are asked which character—Andy or Suzy—should get a thank-you sticker for catching the object (Granny has only one sticker). Experiment 2 is similar except that Granny drops two objects of the same kind (Figure 2b); the consequence of one event is worse than the other because the objects are dropped onto

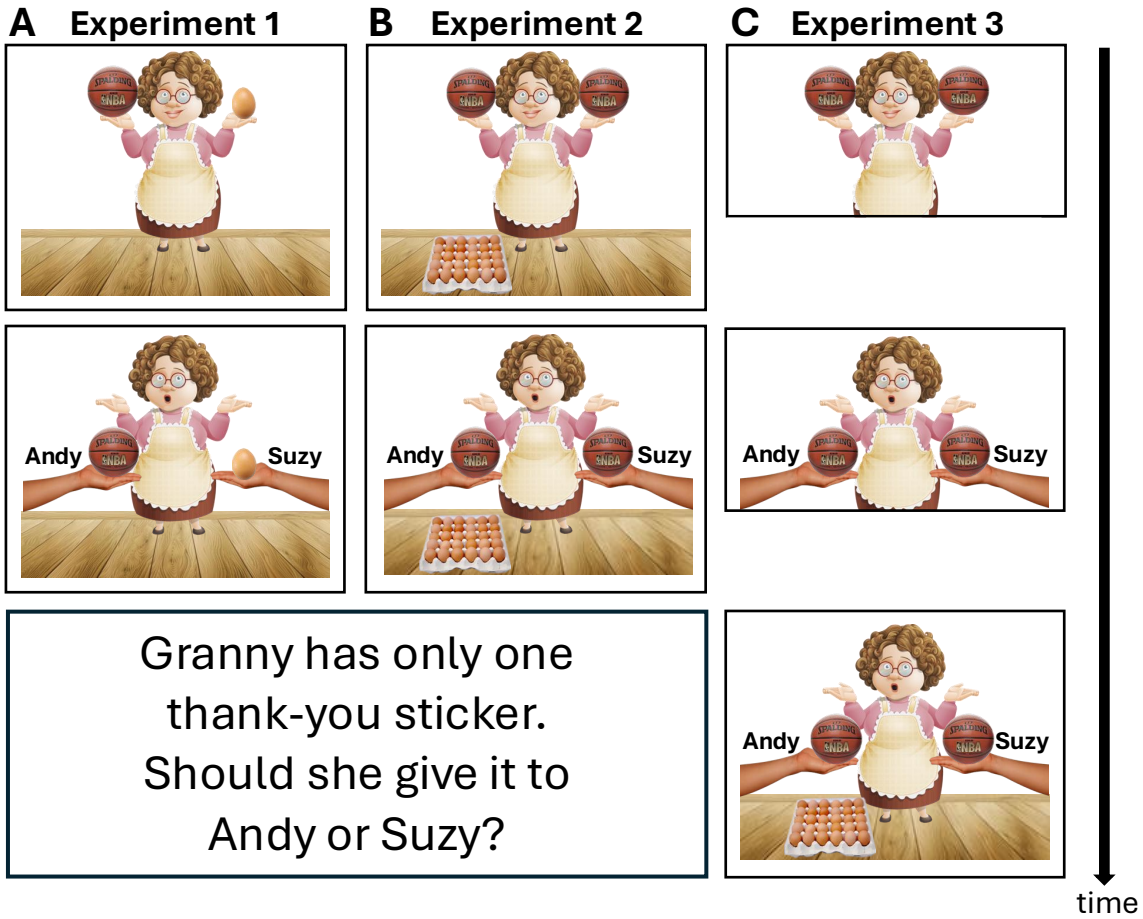

Figure 2

***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? **A** In Experiment 1, Granny drops two different objects. It would have been worse if the egg hadn't been caught than if the basketball hadn't been caught. **B** In Experiment 2, Granny drops two objects of the same kind (e.g., two basketballs). Here, one of the objects would have landed on a carton of eggs and the other on the floor. **C** In Experiment 3, the full scene is revealed only after two objects of the same kind were caught. This means that participants cannot anticipate early on what would happen if Granny dropped the objects.*

different surfaces (e.g., the basketball on a carton of eggs vs. the floor). By engaging in counterfactual thinking (i.e., comparing what would have happened if Andy (or Suzy) hadn't caught the basketball), participants can judge that Andy's action prevented a worse outcome and therefore is more deserving of the sticker than Suzy.

In both of these experiments, however, participants have visual access to the full scene. This means that they could in principle already compute what would happen if Granny 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

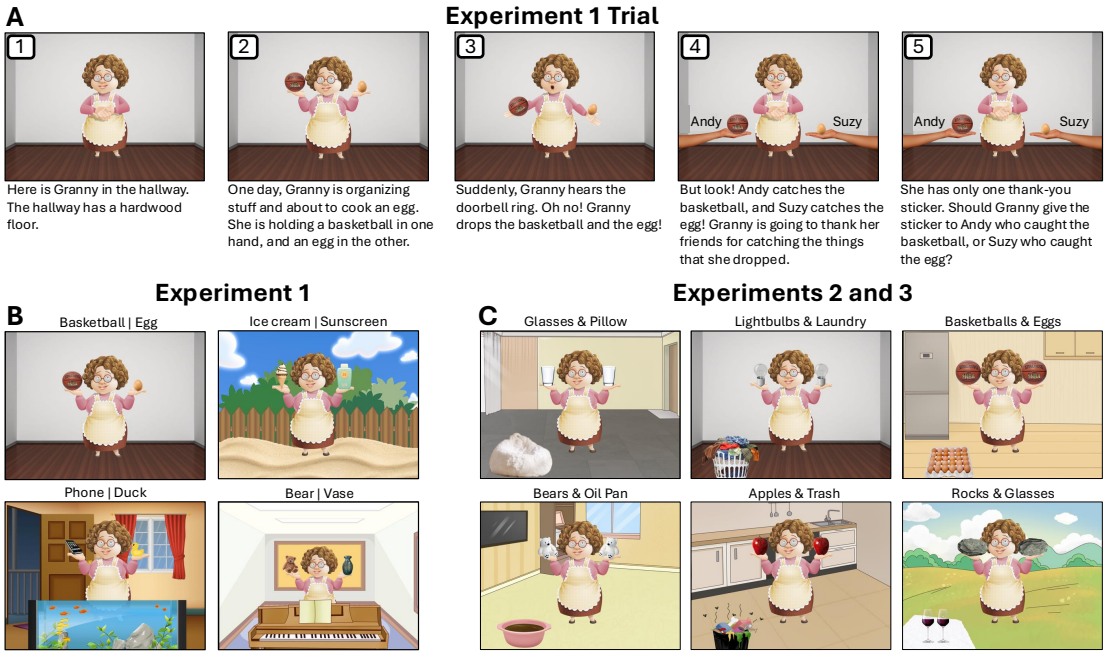

Figure 3

A *Experiment 1 example trial:* ***1*** *Participants first see that Granny is in a hallway. And they are told that the hallway has hardwood floors.* ***2*** *Granny is holding a basketball and egg.* ***3*** *Granny drops the objects after being startled by the doorbell.* ***4*** *The two objects are caught by Andy and Suzy.* ***5*** *Participants are asked who Granny should give the thank-you sticker to.* ***Experiment scenarios****: Each panel in B and C shows a scene (#2) from scenarios that participants saw in each experiment.* ***B*** *In Experiment 1, participants saw four scenarios where Granny drops two different objects.* ***C*** *In Experiments 2 and 3, participants saw six scenarios where Granny drops two of the same objects.*

partial view of the scene, and then reveals the full scene only after the objects have been caught (Figure 2c). Here, counterfactual thinking is necessary to evaluate which outcome would have been worse as one cannot anticipate the possible outcomes in advance.

In what follows, we present results from these three experiments. All experiments were pre-registered. The materials, data, pre-registrations, and analyses are available at `https://github.com/cicl-stanford/counterfactual_development`.

Experiment 1: Counterfactual thinking without counterfactual language

Experiment 1 features scenarios where Granny drops two different objects. Figure 3a shows an example trial and Figure 3b shows the different scenarios. We pre-registered two hypotheses for children. First, that children would be more likely to say that the individual who prevented the worse potential outcome should get the thank-you sticker. Second, that the likelihood of choosing the person who prevented the worse potential outcome from occurring would increase with age. For adults, we pre-registered the hypothesis that they would be more likely to choose the individual who prevented the worse potential outcome.

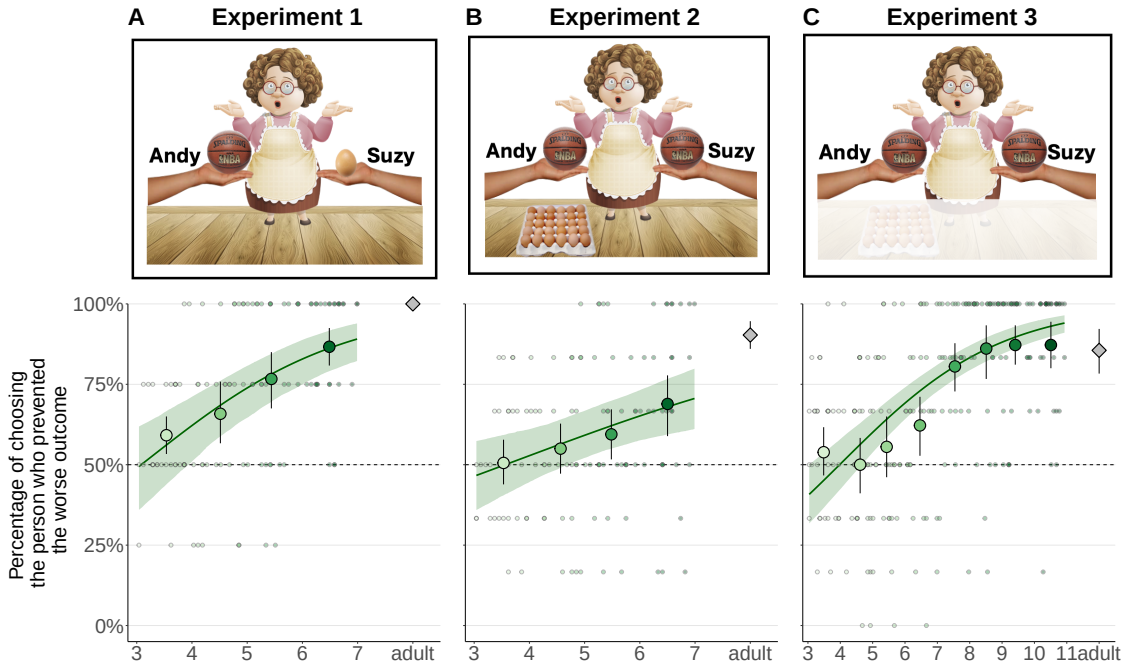

Figure 4

Overall Experimental Results**: Percentage of participants who chose the person who prevented the worse outcome from happening across the three experiments. Lines show the best fit from a Bayesian logistic regression model. The ribbon shows the 95% credible interval for the regression line. Large points show means for each age group. Error bars show 95% bootstrapped confidence intervals. Small points show individual responses averaged across trials. Dashed lines indicate chance performance. In each experiment, there were 30 participants in each age group, except for Experiment 2 which had 31 adults.* ***A *In Experiment 1, two different objects are dropped and prevented from falling onto the ground.* ***B*** *In Experiment 2, two of the same kinds of objects are dropped and prevented from falling onto the ground.* ***C*** *Experiment 3 was similar to Experiment 2, but participants saw where the objects would have landed only after the objects were caught. Initially, only Granny and the two objects were visible. After the objects were caught, the screen zoomed out so that participants could see the entire scene, including the objects at the bottom.*

The results are shown in Figure 4a. For all experiments, we coded responses so that 1 = selecting the person who prevented the worse outcome and 0 = selecting the other person. As predicted, we found that children were more likely to select the individual who prevented the worse outcome from happening as the person who should receive the thank-you sticker, $\beta = 0.73$, 95% credible interval (CrI) = $[0.65, 0.80]$. Older children were more likely to select the individual who prevented the worse outcome from happening than younger children, $\beta = 0.55$, 95% CrI = $[0.34, 0.75]$. The estimated age at which children exceeded chance performance was 3.95 years. Adults chose the individual who prevented the worse outcome from occurring 100% of the time.

These results raise the possibility that children, by around 4 years of age, can already reason about what would've happened if the characters had not caught the objects, and choose the one who prevented the worse outcome from happening. In other words, it is possible that children engaged in counterfactual thinking to decide whom Granny should thank. However, the scenarios used in Experiment 1 always involved two different objects. This leaves open the possibility that instead of simulating which outcome would have been worse, children merely considered what object Granny might like (or even what they themselves like) and chose the person who caught the preferred object (see Figure B1 for relevant empirical data). To address this possibility, Experiment 2 uses a version of the task that involves dropping two identical objects onto two different surfaces.

Experiment 2: Addressing preference-based account

In Experiment 2, we used a version of the task where Granny drops two objects of the same kind (see Figure 3c). This means that children cannot give the correct answer by considering merely which object they themselves or Granny might like better. The pre-registered hypotheses were the same as in Experiment 1. Given that children in Experiment 1 were showing reliably above-chance performance by 4 years of age, we constrained the age range to 3- to 6-year-old children.

The results are shown in Figure 4b. 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: $\beta = 2.44,\ 95\%\ \text{CrI} = [1.49, 3.44]$). Older children were more likely to select the individual who prevented the worse outcome from happening than younger children, $\beta = 0.26,\ 95\%\ \text{CrI} = [0.09, 0.44]$. The estimated age at which children exceeded chance performance was 4.71 years.

Compared to Experiment 1 where two different objects were dropped onto the same surface, children's performance in Experiment 2 was lower. Along with data in Figure B1, this raises the possibility that the results in Experiment 1 could at least be partially explained by children's tendency to choose the person who caught the object that they themselves preferred (or they thought Granny might prefer). This preference-based account, however, does not apply to Experiment 2.[1] Here, 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 concern. In Experiment 2, participants were able to see the whole scene from the very beginning. This means that they were able to simulate the hypothetical of what would happen if Granny dropped one object or the other (before any of this happened). For example, when Granny holds a basketball over a carton of eggs and another basketball over the floor, they can simulate that dropping the basketball over the egg carton would be worse than dropping the basketball over the floor. They can then compare the outcome of these hypothetical simulations to what actually happened, and make the judgment that catching the basketball over the egg carton is more deserving of a

[1] While it is in principle possible that some children preferred one type of surface over another, it is highly unlikely that these preferences would be systematic.

reward. Experiment 3 addresses this concern.

Experiment 3: Genuine counterfactual thinking

Experiment 3 is similar to Experiment 2, except that the full scene is revealed only *after* the objects were caught (see Figure 4c). This subtle difference has a powerful effect: it renders hypothetical thinking ineffective by making it impossible to anticipate what would happen until the outcomes are revealed. As such, this experiment provides a stronger test for children's ability to think counterfactually. Our pre-registered hypotheses were the same as in Experiments 1 and 2. This time, we expanded the age range from 3- to 10-year-old children because we anticipated that children might find this version of the task more challenging.

The results are shown in Figure 4c. 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: $\beta = 1.19,\ 95\%\ \text{CrI} = [0.95, 1.45]$, adults: $\beta = 2.78,\ 95\%\ \text{CrI} = [0.74, 4.99]$). Older children were more likely to select the individual who prevented the worse outcome from happening than younger children, $\beta = 0.40,\ 95\%\ \text{CrI} = [0.31, 0.50]$. The estimated age at which children exceeded chance performance was 5.02 years.[2]

In sum, even when key information about where the two objects might land was unavailable until the end of the scenario—effectively preventing children from using hypothetical reasoning to answer the question—the results were comparable to Experiment 2; children showed above-chance responses by 5 years of age. Additionally, by testing a wider age range that includes older children, we were able to observe a substantial increase in accuracy, nearing ceiling (and adult performance) by age 7–8.

General Discussion

Counterfactual thinking is a fundamental cognitive capacity. Understanding when children begin to engage in counterfactual thinking is important because it helps us better understand how different cognitive capacities—such as causal reasoning, experiencing regret, learning from mistakes, and social and moral reasoning—are related to one another. Yet, prior studies on how and when children develop this ability have yielded mixed results, due to at least two reasons: first, they use complex counterfactual language (which may have underestimated children's competence); second, they use tasks where success can be explained by conditional or hypothetical thinking rather than genuine counterfactual thinking. The current work presents a novel paradigm that doesn't use counterfactual language and that teases apart hypothetical and counterfactual thinking.

To identify young children's ability to think counterfactually, our paradigm leverages their intuitive understanding of the physical and social world. The scenarios involve simple physical events and actions: objects being dropped and being caught. While dropping an object is generally a negative event, one event was deliberately designed to be worse than

[2]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 is due to the high performance of older children. We would expect a similar shift for Experiments 1 and 2, if the same age range was tested as in Experiment 3.

the other (i.e., dropping a basketball onto a carton of eggs is worse than dropping it onto the floor). Given that young children have plenty of experience dropping objects themselves (or observing such events), and with adults expressing their emotions about that, our scenarios were likely easy to understand even for the youngest participants. Furthermore, saying "thank you" or receiving a sticker for doing something good is a social act that children learn early on (e.g., Noles & McDermott, 2023; Vaish & Savell, 2022). Our task leveraged this familiarity to ask children to choose which one of the two characters should get a thank-you sticker. Despite its apparent simplicity, choosing whom to thank in our task requires simulating the outcome of each object dropping all the way to figure out who prevented the worse outcome from happening. By design, deciding who should receive the thank-you sticker probes counterfactual thinking without requiring counterfactual language, such as explicitly asking children what would have happened if an object hadn't been caught.

Another key feature of our paradigm is that it teases apart counterfactual thinking from hypothetical thinking. Some prior work on counterfactual thinking has used counterfactual questions that can be answered by recalling an event that was observed earlier in the task (e.g., Nyhout & Ganea, 2019). In contrast, in our task, children never see an object actually hitting the floor—they have to mentally simulate what the counterfactual outcome would have been. Using different versions of this task across three experiments that get incrementally closer to genuine counterfactual thinking, we were able to estimate the age at which children begin to think counterfactually: around age 5.

Of course, pinning down the *exact* age at which counterfactual thinking develops may not be a feasible scientific goal; no experimental task is completely free from task demands or other cognitive prerequisites. For instance, our task still requires some language to understand the scenarios. A completely non-verbal task for assessing counterfactual thinking does not exist yet; even in prior work that showed counterfactual simulations in adults using their eye movements (see Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017), at least some verbal instructions were necessary to guide participants. Yet, such non-verbal measures may provide useful ways to further reduce language demands. One might also wonder whether our task has fully ruled out the possibility that participants succeed via hypothetical, rather than counterfactual thinking. While participants cannot run hypothetical simulations before the objects were dropped in Experiment 3, they could in principle run such simulations after knowing what happened. Note however that the distinction between hypothetical and counterfactual simulation becomes blurry here; if we knew everything that happened in advance, then a hypothetical simulation would yield the same result as a counterfactual simulation (because there is no benefit of hindsight anymore). These two forms of thinking only come apart when there is some uncertainty about the future that gets resolved based on what happened (Gerstenberg, 2022). Our experiment was designed precisely to create this resolution point. In doing so, it ensures that participants were not just imagining possible futures, but were actively contrasting what in fact occurred with what would have happened otherwise—the hallmark of counterfactual thought.

By offering more precise age estimates, we can better understand what cognitive, physical, and social capacities can be acquired without counterfactual thinking, and how counterfactual thinking might help in further developing these and other related capacities. Our results suggest that the ability to engage in counterfactual thinking emerges by around

five years of age. While this initial estimate is based on U.S. children tested asynchronously through an online sample, the real strength of our work is in offering an intuitive task that can easily be adapted and administered in other languages across different platforms and cultural contexts. In this sense, our work opens up a range of empirical questions about the consistency as well as the variability in the development of counterfactual thinking.

So what does all of this imply? If it is indeed the case that children only develop the capacity for counterfactual thinking 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. And this suggests that counterfactual thinking may not be the early foundation on which much of higher cognition rests, but rather a later-developing capacity, one that perhaps reshapes and deepens abilities that are already in place. Nonetheless, counterfactual thinking may play a transformative role: once in place, counterfactual thought enables richer forms of explanation, planning, and moral evaluation that go beyond what earlier-developing systems can achieve.

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. Much like how the Sally-Anne task provided a novel way to study the development of children's understanding of mental states (Baron-Cohen, Leslie, & Frith, 1985; Dennett, 1978; Wimmer & Perner, 1983)—one that revolutionized research on theory of mind—our task has the potential to change the empirical landscape surrounding the development of counterfactual thinking.

Materials and Methods

Experiment 1: Counterfactual thinking without counterfactual language

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. To be included, participants needed to pass the warmup trial. Since our task involved a number of friends' names and objects across trials, we wanted to ensure that participants could track this. To accommodate younger participants, we allowed children to respond by pointing, with parents verbally reporting their child's answer. 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. 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 = 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 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 3b. 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 3a. 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. 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

In all three experiments, we fit separate Bayesian logistic regression models for children and adults with random intercepts for both participants and trials. 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). We report the means of the posterior distributions together with the 95% credible intervals. Because we ran logistic regression models,

the quantities are presented on the log odds scale. Results for each trial can be seen in Figure A1.

Experiment 2: Addressing preference-based account

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; *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 3c.

Procedure

The procedure and design were the same as in Experiment 1. An example trial is shown in Figure C1.

Results

Results for each trial can be seen in Figure C2.

Experiment 3: Genuine counterfactual thinking

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 and design were the same as in Experiment 2 except that it was revealed only after the objects were caught, where they would have fallen. An example trial is shown in Figure D1.

Results

Results for each trial can be seen in Figure D2.

Author contributions

Conceptualization: DR, SZ, SB, HG & TG; Methodology: DR, SZ, SB, HG & TG; Software: DR, SZ, TG; Validation: DR, TG; Formal Analysis: DR, TG; Investigation: DR, SZ; Data Curation: DR, SZ; Writing—Original Draft: DR; Writing—Review & Editing: DR, HG, TG; Visualization: DR, TG; Supervision: DR, HG, TG; Project Administration: DR, HG, TG; Funding Acquisition: TG

Acknowledgments

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

Amsel, E., & Smalley, J. D. (2014). Beyond really and truly: Children's counterfactual thinking about pretend and possible worlds. In *Children's reasoning and the mind* (pp. 121–147). Psychology Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Beck, S. R., & Crilly, M. (2009). Is understanding regret dependent on developments in counterfactual thinking? *British Journal of Developmental Psychology*, *27*(2), 505–510.

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.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.-R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science*, *304*(5674), 1167–1170.

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.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain sciences*, *1*(4), 568–570.

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

Guttentag, R., & Ferrell, J. (2004). Reality compared with its alternatives: age differences in judgments of regret and relief. *Developmental Psychology*, *40*(5), 764.

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

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.

Jones, A. K., Nelson, N. L., Gautam, S., & Redshaw, J. (2025). Children infer counterfactual information from others' facial expressions. *Journal of Experimental Child Psychology*, *259*, 106306.

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.

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78. Retrieved from http://dx.doi.org/10.1016/j.tics.2019.11.004 doi: 10.1016/j.tics.2019.11.004

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., Feeney, A., & Beck, S. R. (2020). Regret and decision-making: A developmental perspective. *Current Directions in Psychological Science*, *29*(4), 346–350.

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

McCormack, T., O'Connor, E., Beck, S., & Feeney, A. (2016). The development of regret and relief about the outcomes of risky decisions. *Journal of Experimental Child*

Psychology, *148*, 1–19. doi: 10.1016/j.jecp.2016.02.008

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

Noles, N. S., & McDermott, C. H. (2023). Children's understanding of gratitude, generosity, and reciprocity. *Cognitive Development*, *66*, 101322.

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

O'Connor, E., McCormack, T., & Feeney, A. (2012). The development of regret. *Journal of experimental child psychology*, *111*(1), 120–127.

Ozturk, O., & Papafragou, A. (2015). The acquisition of epistemic modality: From semantic meaning to pragmatic interpretation. *Language learning and development*, *11*(3), 191–214.

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.

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., & Perner, J. (2012). When the alternative would have been better: Counterfactual reasoning and the emergence of regret. *Cognition & emotion*, *26*(5), 800–819.

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., 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.

Roese, N. J. (1994). The functional basis of counterfactual thinking. *Journal of personality and Social Psychology*, *66*(5), 805.

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

Shtulman, A., & Phillips, J. (2018). Differentiating "could" from "should": Developmental changes in modal cognition. *Journal of Experimental Child Psychology*, *165*, 161–182.

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.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychol-*

ogy, *66*(1), 223–247. Retrieved from http://dx.doi.org/10.1146/annurev-psych-010814-015135 doi: 10.1146/annurev-psych-010814-015135

Smallman, R., & McCulloch, K. C. (2012). Learning from yesterday's mistakes to fix tomorrow's problems: When functional counterfactual thinking and psychological distance collide. *European journal of social psychology*, *42*(3), 383–390.

Stanley, M. L., Cabeza, R., Smallman, R., & De Brigard, F. (2021). Memory and counterfactual simulations for past wrongdoings foster moral learning and improvement. *Cognitive Science*, *45*(6), e13007.

Summerville, A. (2011). Counterfactual seeking: The scenic overlook of the road not taken. *Personality and Social Psychology Bulletin*, *37*(11), 1522–1533.

Vaish, A., & Savell, S. (2022). Young children value recipients who display gratitude. *Developmental Psychology*, *58*(4), 680.

Weisberg, D. P., & Beck, S. R. (2012). The development of children's regret and relief. *Cognition & emotion*, *26*(5), 820–835.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

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.

Appendix A
Experiment 1

Trial results

The results for each of the four trials are shown in Figure A1. As can be seen, they are largely consistent across trials.

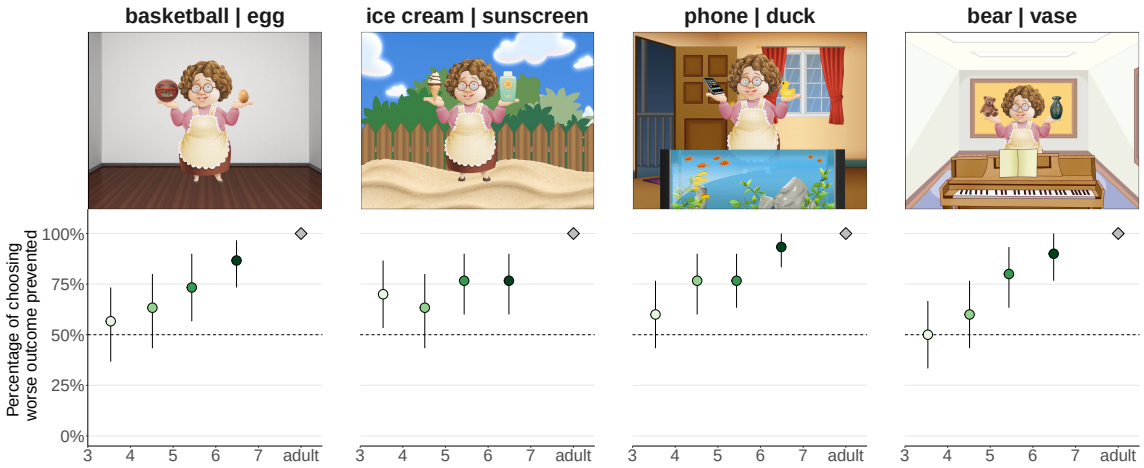

Figure A1

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. The pattern of performance was similar across the four different scenarios.*

Appendix B
Experiment 1: Which object do you like? Which object does Granny like?

In this experiment, we asked whether children might select who to thank by thinking about what object they like or what object Granny likes.

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*: 57 female, 63 male; *language*: 117 English, 3 no response/other). Each age group included 30 participants. Families received $5.

Materials

The materials were the same as in Experiment 1.

Procedure

Children completed two blocks that included four trials each. In each block, Granny is holding the same pairs of objects as in Experiment 1. The setup within each block is

identical, featuring the same content but what question children were asked to answer differed between each block. In one block, children were asked "Which one do you like more?'. In the other, they were asked "Which one do you think Granny likes more?". In both blocks, we randomized on what side each object was shown. The order of test trials within each block was also randomized. Lastly, the presentation order of the two blocks was counterbalanced.

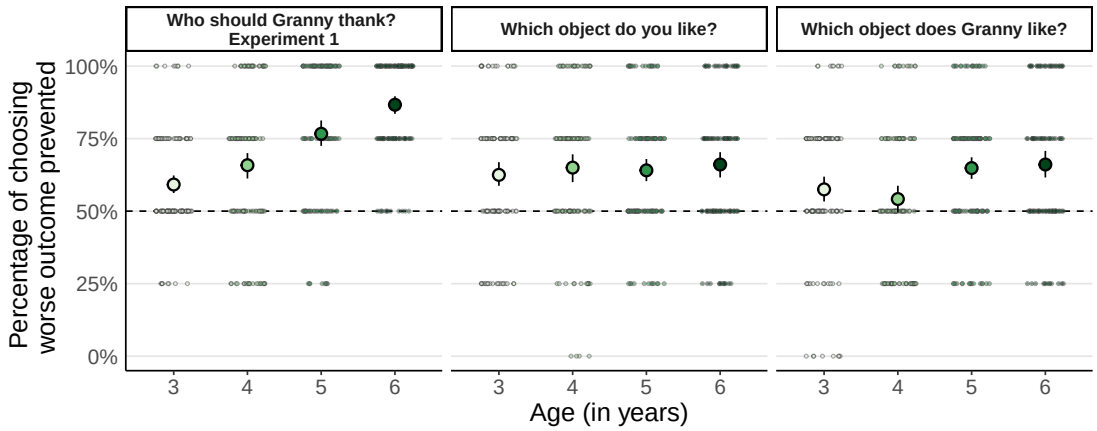

Figure B1

Experiment 1: Which object do you like? Which object does Granny like?: *Percentage of participants who chose the person who prevented the worse outcome from happening. Experiment 1 results are shown on the left for comparison. 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.*

Results

The results are shown in Figure B1. As can be seen, children's performance is similar in the preference tasks to the thank-you sticker task, but the age trend is stronger for the thank-you sticker task.

Appendix C
Experiment 2

Example trial

Figure C1 shows an example trial from Experiment 2.

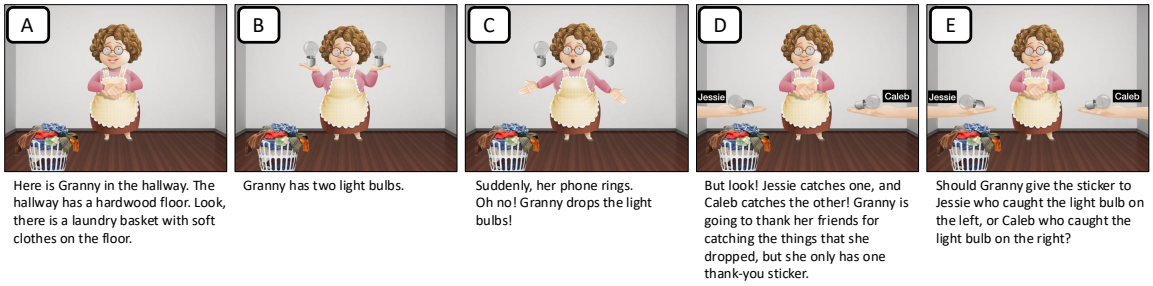

Figure C1

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.*

Trial results

The results for each of the six trials are shown in Figure C2. As can be seen, they are largely consistent across trials.

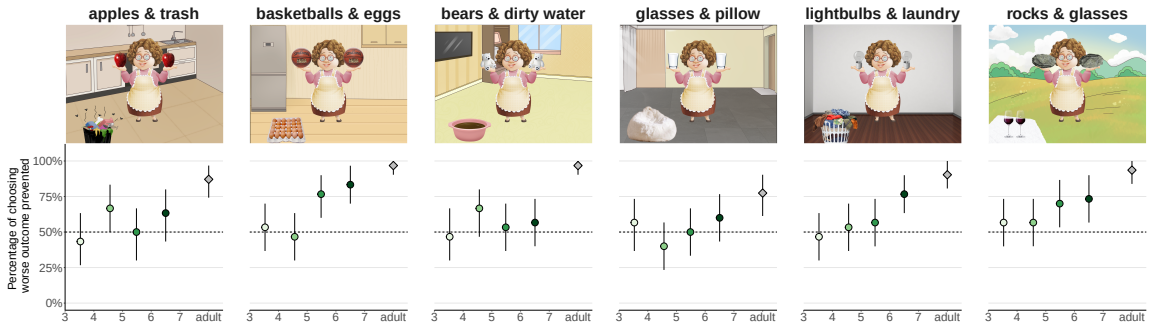

Figure C2

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. There was some variability in performance across the different trials. Notably, even adult participants had mixed intuitions about the glasses & pillow trial.*

Appendix D
Experiment 3

Example trial

Figure D1 shows an an example trial from Experiment 3.

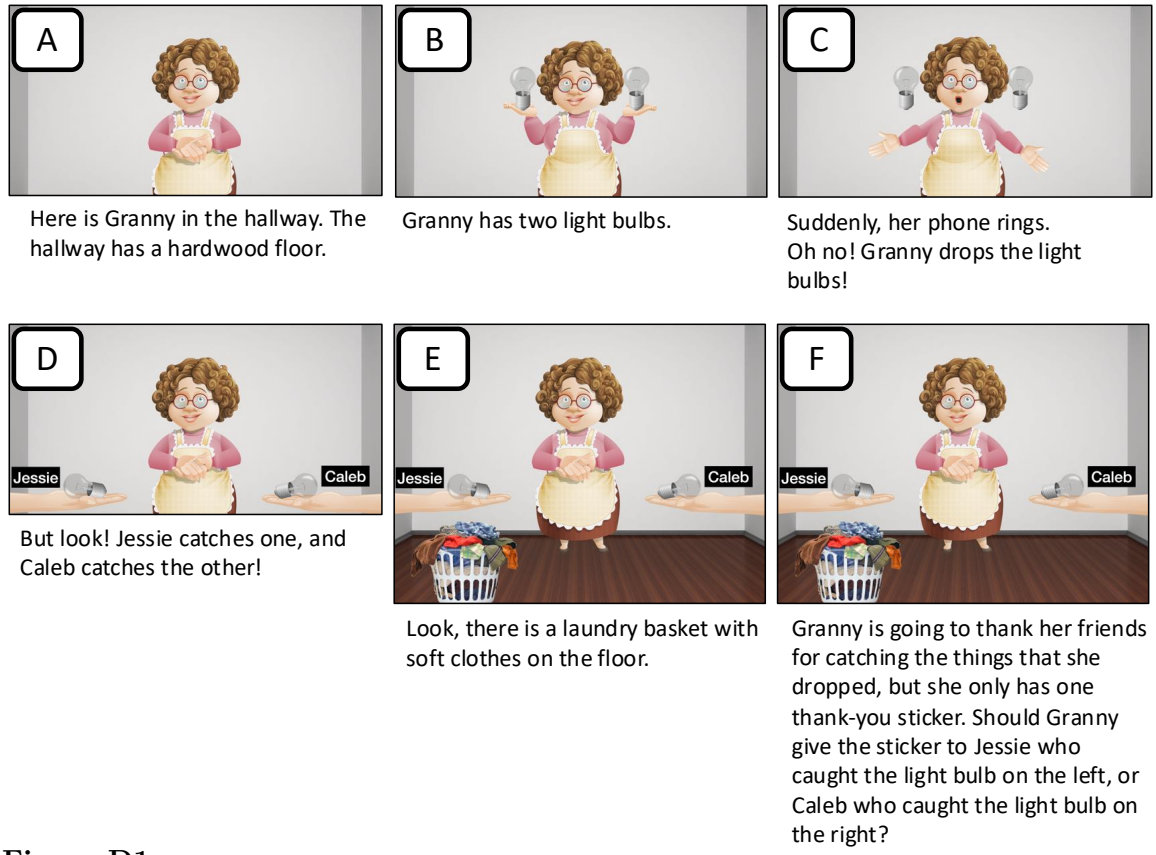

Figure D1

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.*

Trial results

The results for each of the six trials are shown below. As can be seen in Figure D2, 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.

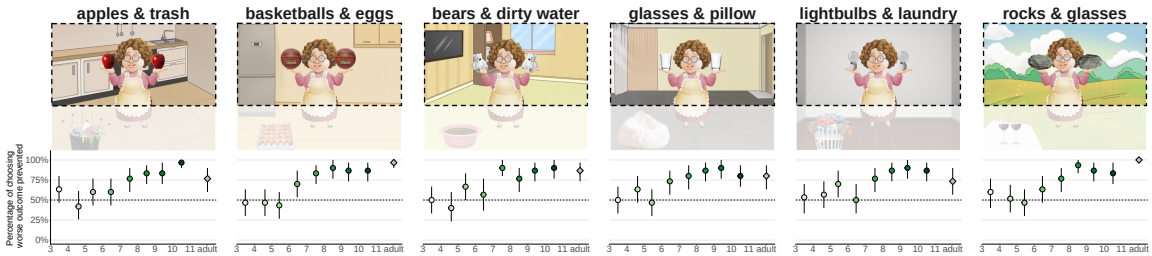

Figure D2

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.*