The trajectory of counterfactual simulation in development

Jonathan F. Kominsky[a], Tobias Gerstenberg[b], Madeline Pelz[c], Mark Sheskin[d,e], Henrik Singmann[f], Laura Schulz[c], & Frank C. Keil[e]

[a]Rutgers University – Newark, [b]Stanford University, [c]Massachusetts Institute of Technology, [d]Minerva Schools at KGI, [e]Yale University, [f]University of Warwick

Corresponding author: jonathan.f.kominsky@gmail.com; Stimulus materials, data, and analysis code and output for all experiments are available at our OSF repository: https://osf.io/5jw6y/

Abstract

Young children often struggle to answer the question "what would have happened?", particularly in cases where the adult-like 'correct' answer has the same outcome as the event that actually occurred. Previous work has assumed that children fail because they cannot engage in accurate counterfactual simulations. Children have trouble considering what to change and what to keep fixed when comparing counterfactual alternatives to reality. However, most developmental studies on counterfactual reasoning have relied on binary yes/no responses to counterfactual questions about complex narratives, and so have only been able to document *when* these failures occur but not *why* and *how*. Here, we investigate counterfactual reasoning in a domain in which specific counterfactual possibilities are very concrete: Simple collision interactions. In Experiment 1, we show that 5-10-year-old children (recruited from schools and museums in Connecticut) succeed in making predictions but struggle to answer binary counterfactual questions. In Experiment 2, we use a multiple-choice method to allow children to select a *specific* counterfactual possibility. We find evidence that 4-6-year-old children (recruited online from across the USA) do conduct counterfactual simulations, but the counterfactual possibilities younger children consider differ from adult-like reasoning in systematic ways. Experiment 3 provides further evidence that young children engage in simulation rather than using a simpler visual matching strategy. Together, these experiments show that the developmental changes in counterfactual reasoning are not simply a matter of whether children engage in counterfactual simulation, but also how they do so.

Keywords. *Counterfactual reasoning; Mental simulation; Cognitive development; Intuitive physics; Multinomial Process Tree models*

Introduction

When considering whether one event caused another, adults do not merely consider what actually happened. Rather, we think about what could, or would, or should have happened had the causal event been altered in some way (Byrne, 2016; Lewis, 1973). Counterfactual reasoning is a central aspect of adult causal cognition. There is more to causality than actuality – what would have happened in relevant counterfactual possibilities radically affects causal judgments about agents, objects, and events (Gerstenberg, Goodman, Lagnado, & Tenenbaum, submitted; Icard, Kominsky, & Knobe, 2017; Kominsky & Phillips, 2019; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Phillips, Luguri, & Knobe, 2015). The ability to consider counterfactual possibilities underlies the emotions of regret and relief (Beck, Weisberg, Burns, & Riggs, 2014; McCormack, O'Connor, Beck, & Feeney, 2016; O'Connor, McCormack, & Feeney, 2012), and it has been argued that counterfactual reasoning is critical for learning, inference, and decision-making (Pearl, 2000; Roese, 1997). In fact, counterfactual reasoning is so central to mature causal cognition that adults engage in it spontaneously (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; McEleney & Byrne, 2006).

One of the essential properties of counterfactual reasoning is that it often involves what can broadly be called *simulation*. Counterfactual reasoning operates on a mental causal model that represents what actually happened, but also supports simulating what would have happened if an event of interest had been different. This "episodic simulation" (Kahneman & Tversky, 1982; Mahr, 2020) allows reasoners to preserve the causal structure and relationships of the original scenario while evaluating the effect of altering a particular cause (Sloman & Lagnado, 2005).

Despite the centrality of counterfactual simulation to adult causal reasoning, the emergence and development of this ability in childhood is not well-understood and still very much under debate. The earliest body of work on the matter found that children were incapable of counterfactual reasoning until around age 12 (Inhelder & Piaget, 1958). Later work argued that children could answer counterfactual questions as early as three years of

age. For example, in a classic study by Harris, German, and Mills (1996), children were presented with scenarios like this: "One day, the floor is clean. But guess what? Carol comes home and she doesn't take her shoes off. She comes inside and makes the floor all dirty with her shoes." (Harris et al., 1996, p. 238). When asked "What if Carol had taken her shoes off — would the floor be dirty?", even 3–year-olds responded that the floor would be clean. However, later work suggests that this early success may have been overstated. If presented with an *over-determined* version of the scenario, in which two people who walk across the floor in dirty shoes (e.g., "Carol *and Max* come home, don't take their shoes off, and make the floor dirty with their shoes. What if Carol had taken her shoes off?"), children fail: While adults say the floor would still be dirty (as only one person took their shoes off), 5–6-year-olds systematically said that the floor would be clean, while 9–10-year-olds were at chance, and not until 14 years of age did performance reach adult-like levels (Rafetseder, Schwitalla, & Perner, 2013).

Since then, the literature has provided oscillating estimates of when the ability to engage in counterfactual reasoning emerges. Using dynamic events as stimuli, some authors have found that children can correctly answer counterfactual questions in over-determined cases at age 6 (Beck & Guthrie, 2011; McCormack, Ho, Gribben, O'Connor, & Hoerl, 2018) or even age 4 (Nyhout & Ganea, 2019). Even using similar narratives to Rafetseder et al. (2013), small modifications to the causal structure of the events allowed children to succeed as young as age 6–8 (Nyhout, Henke, & Ganea, 2019). In a related literature, researchers have found that children between ages 6 and 10 experience emotions like regret and relief that require counterfactual processing (Ferrell, Guttentag, & Gredlein, 2009; O'Connor et al., 2012; Payir & Guttentag, 2019). Again, there is no consensus about when in that age window counterfactual emotions emerge.

There is disagreement not only about when children succeed at counterfactual reasoning, but also why they fail. The earliest view, offered by Piaget, was that they simply lacked the capacity. Under this view, counterfactual reasoning requires children to reach the stage of "formal operations", the stage of reasoning that allows for the manipulation

of abstract information (Inhelder & Piaget, 1958). Others, noting that young children have no difficulty with hypothetical reasoning about the future or with pretense (Atance & O'Neill, 2005; Buchsbaum, Bridgers, Skolnick Weisberg, & Gopnik, 2012), have argued that children are able to imagine possible situations in general, but fail specifically when asked counterfactual questions. What makes counterfactual reasoning special is that it requires simultaneously representing events as they actually occurred as well how they would have played out had something about the past been different (Beck & Riggs, 2014; Beck et al., 2014). This imposes several cognitive challenges that place substantial demands on children's executive function abilities: children have to keep multiple possibilities in mind at the same time (Beck, Robinson, Carroll, & Apperly, 2006; Carey, Leahy, Redshaw, & Suddendorf, 2020; Rafetseder et al., 2013) and they have to inhibit what actually happened when considering counterfactual alternatives (Beck, Riggs, & Gorniak, 2009; Carlson, White, & Davis-Unger, 2014).

Under these views, children answer counterfactual questions by relying on alternative strategies rather than engaging in counterfactual simulation. For example, Rafetseder et al. (2013) argued that children use "basic conditional reasoning" (BCR) to answer counterfactual questions. This BCR strategy differs from true counterfactual reasoning in that it does not try to preserve the features of the event as it actually occurred. Instead, BCR allows everything about the event that can be changed to change (Leahy, Rafetseder, & Perner, 2014). In over-determined cases, this means that young children believe the outcome would be different because they ignore the fact that a second cause was actually present, while adults preserve the state of any cause other than the one specified by the counterfactual question.

However, there is another way in which children could fail which previous work has not considered: Perhaps children do engage in counterfactual simulation, but they systematically consider different alternatives than we would as adults. That is, when we say adults answer these questions 'correctly', we mean that their answers are compatible with what we consider to be the appropriate counterfactual simulation given the conditions posited in

the question. For example, in the case of the dirty shoes, when there are two people who walk across the dirty floor and adults are asked what would happen if one of them had taken their shoes off, the 'correct' answer, that the floor would still be dirty, presumes that the counterfactual we simulate is one in which the person not mentioned in the question leaves their shoes on (Sloman & Lagnado, 2005). Younger children, who systematically say that the floor would be clean, could do so on the basis of an episodic simulation in which *both* people take their shoes off, and so arrive at the 'wrong' answer while still engaging in counterfactual simulation.

We propose that children engage in counterfactual simulation (possibly at an earlier age than prior research suggests), but they consider different counterfactual possibilities than adults. Importantly, previous work has largely relied on binary forced-choice questions, and so could only determine whether children arrive at the 'correct' or adult-like answer (e.g., the failures documented by Piaget and others), not how or why they fail. As a result, this 'different simulation' explanation makes the same prediction about previous results as non-simulation explanations (like BCR), because the measures used cannot distinguish the two. To determine what children might be simulating when asked counterfactual questions, a different method is needed that does not rely on binary choices.

This project had two goals. First, to test the hypothesis that children answer counterfactual questions incorrectly because they do engage in counterfactual simulation, but simulate different possibilities than adults. Second, to investigate not only whether children systematically simulate different possibilities, but if they do, to provide an initial investigation of the ways in which their simulations might differ from those of adults.

The present experiments

In order to examine which specific counterfactual possibilities children consider, we depart from the narrative studies that have been used in most prior work. Narrative stimuli add a great deal of memory load and room for influence from idiosyncratic knowledge. The ideal stimuli would be a causal event that children understand nearly effortlessly, that

minimizes memory load when answering a counterfactual question, and that can be systematically manipulated to examine not just whether children are simulating counterfactual alternatives, but which specific alternatives they consider.

To that end, our experiments use simple Newtonian collision events, which are effortlessly and automatically understood as early as 6 months of age (Kominsky et al., 2017; Leslie & Keeble, 1987; Saxe & Carey, 2006). Similar displays have recently been used to study counterfactual simulation in adults (Gerstenberg et al., 2017), as well as the influence of counterfactual reasoning on causal (Gerstenberg & Icard, 2019).

However, such displays have never been used in developmental studies of counterfactual reasoning. Therefore, Experiments 1a and 1b first seek to replicate previously observed patterns of successes and failures in children with these new stimuli. Experiment 1a tests whether children are able to correctly answer binary counterfactual questions when the outcome would have been different in the relevant counterfactual situation, and when it would have been the same. Experiment 1b tests whether children succeed at making *future hypothetical* judgments (i.e. predictions) about the same events.

Experiment 2 then uses a novel four-alternative forced-choice paradigm inspired in part by the methods of Rafetseder and Perner (2018), in which children answer a counterfactual question by choosing one out of four specific counterfactual alternatives. By constructing these alternatives in a systematic way, we are able to investigate not only whether children choose the normatively 'correct' answer, but also whether they are systematic in their wrong answers. If children fail to simulate altogether, one might expect them to pick randomly among the options they are offered. However, if they are simulating differently from adults, then they should systematically prefer certain alternatives over others. To foreshadow our results, we find that children systematically prefer certain counterfactual possibilities over others. In Experiment 3, we present evidence against an alternative hypothesis for the observed pattern of results in Experiment 2 – that children simply choose alternatives that most closely visually match what actually happened.

Experiments 1a and 1b: Forced-choice counterfactual judgments and predictions

The goal of Experiment 1a was to replicate previous findings in the developmental literature on counterfactual reasoning in the novel domain of simple collision events. In particular, the goal was to determine whether children would succeed at answering counterfactual questions about the outcome of simple collision events when the outcome was singly-determined (i.e., the outcome in the counterfactual situation would have been different), but would fail when the outcome was over-determined (i.e., the outcome in the counterfactual situation would have been the same).

Experiment 1a: Counterfactual simulation

Methods.

Participants. We planned to run 40 children in each age group (20 in each of two conditions), and continued collecting data until we had reached that target, replacing any participants that were excluded. 40 5-6-year-olds (15 female), 40 7-8-year-olds (15 female), and 40 9-10-year- olds (18 female) participated in Experiment 1a, recruited from local schools and children's museums in southern Connecticut. In addition, 10 5-6-year-olds (5 female), 3 7-8-year-olds (2 female) and 1 (male) 9-10-year-old participated but were excluded from analyses based on predetermined exclusion criteria (see below).

Stimuli and procedure. Experiments 1a and 1b were approved by the Yale University IRB under protocol # 1311013027, "Cognitive and metacognitive development". We constructed simple animations modeled on those used by Gerstenberg, Goodman, Lagnado, and Tenenbaum (2015) (see Fig. 1a. Videos of the animations can be found at `https://osf.io/5jw6y/`). In these animations, there are two balls, A and E, a red area that was described as a "goal", and black walls on either side of the goal. The survey was administered via Qualtrics (Qualtrics, 2005) and presented on an iPad.

All participants first saw two training items in counterbalanced order. In one training item, ball A hit ball E, which then bounced off the boundary above the goal and off the field

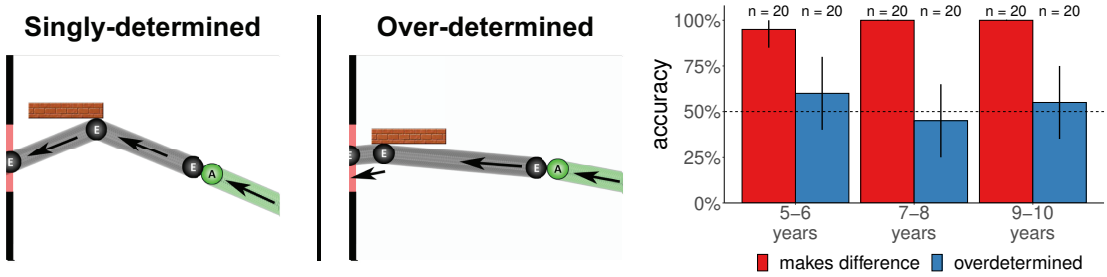
entirely. In the other training item, ball A hit ball E, which then went into the goal directly. Following each training trial, participants were asked two questions: "Before ball A hit ball E, was ball E moving or sitting still?", and "Did ball E go into the goal?" Participants could either respond verbally and the experimenter recorded their answer, or they could select one of the response options ("yes" or "no") on the iPad directly. If participants answered either question incorrectly on one of the training trials, they were shown that training animation a second time and asked again. No child answered incorrectly on the second attempt. For all questions at both training and test, participants did not see the event or the field while answering the question.

Participants then saw one of two test trials, between subjects. In the "singly-determined" condition, the animation was almost identical to the training item in which ball E bounced off the wall above the goal, except that there was an added "brick wall" (see Fig. 1a) that ball E bounced off of before it went into the goal. In this clip, ball E would have missed the goal if the brick wall had been absent. In the "over-determined" condition, the animation was almost identical to the training item in which the ball went into the goal, except that the ball bounced off the brick wall before going into the goal. In this clip, ball E would have gone through the goal even if the brick wall had been absent.

Following the test trial, participants were asked the same two questions as in the training trials. If children answered either question incorrectly, they were not corrected but their data were excluded. Then, children were asked the critical test question: "What if the brick wall had not been there? Would ball E have gone into the goal?". Participants once again replied by selecting "yes" or "no".

Results. Fig. 1b shows the results. A simple inspection of this figure gives a clear sense of the results, which were similar across all age groups: Near-perfect performance on cases in which the brick wall made a difference (where the correct answer is that ball E would not have gone into the goal), but only roughly 50% accuracy for over-determined events (where the correct answer is that ball E would still have gone into the goal).

This impression was verified with a binary logistic regression with age group and



(a) Diagrams of the two test clips (b) Results

Figure 1. **Experiment 1a**: (a) Diagrams of the test trials. In the singly-determined event (left), the brick wall altered ball E's trajectory such that it went into the goal. In the over-determined condition (right), ball E also deflects of the wall, but would have gone into the goal regardless. (b) Proportion of accurate responses to the counterfactual question separated by age group and condition (whether the brick wall made a difference to ball E's going through the gate (red), or whether the outcome was over-determined (blue)). The dashed line at 50% represents chance responding. Error bars are 95% bootstrapped confidence intervals.

condition as factors. This analysis revealed a main effect of condition, $\beta = 2.54, p = .02$, but no detectable effect of age group and no interactions, ps $> .9$. As children demonstrated nearly uniform perfect performance in the singly-determined condition (one incorrect answer in total), no further analyses were conducted for this condition. For the over-determined condition, a logistic regression with age group also showed no effect of age ($p > .3$) and no significant intercept ($p = .37$). We conducted a binomial exact test of performance in the over-determined condition collapsed across age, which found no significant difference from chance responding (chance being 50%), $p = .7$.

Experiment 1b: Hypothetical simulation

Methods.

Participants. This study was stopped early due to the fact that all children responded correctly.[1] Our final sample sizes were therefore 21 5-6-year-olds (10 female) and

[1] We acknowledge that this is an atypical decision, since we did not use a preset 'stopping rule' but rather stopped data collection arbitrarily. However, after the unprecedented experience of receiving the *exact* same response from 47 participants, we decided that further data collection was not justifiable.

26 7-8-year-olds (14 female) recruited from the same populations as Experiment 1a. In addition, 4 5-6-year-olds (2 female) and 1 (male) 7-8-year-old were excluded based on predetermined exclusion criteria (see below).

Stimuli and Procedure. The goal of this study was to look at children's hypothetical judgments about the same cases tested in Experiment 1a. The stimuli were similar to Experiment 1a with the following differences: Participants first saw four training trials in random order: Two in which ball E went into the goal and two in which it missed the goal. First, children saw an animation where ball A struck ball E, and ball E moved approximately halfway from its starting position to the left edge of the display (where the wall and goal are located). At this point the animation froze and a large "pause" icon appeared (that didn't obstruct either of the balls). Children were then asked, "If ball E keeps going, will it go into the goal?" Children could respond "yes" or "no". For the training trials, children then saw the rest of the animation. If children made incorrect predictions on at least two of these items, they were excluded from analyses on the basis that they did not understand the task.

Following training, children saw two test trials, a "singly-determined" trial and an "overdetermined" trial in counterbalanced order. The test trials were identical to those used in Experiment 1a, with two exceptions: First, the brick wall was not visible (i.e., identical to Experiment 1a's training trials). Second, the animation paused on the frame in which the ball would have collided with the brick wall in Experiment 1a (participants had no way of knowing this). Participants were then asked the same question as in the training items, but were not shown the end of the animation. Note that the predictions that children are asked to make in Experiment 1b are identical to the counterfactual simulation that is required to answer what would have happened without the brick wall in Experiment 1a.

Results. Every single child who passed the training provided correct answers to both test questions (21/21 5-6-year-olds and 26/26 7-8-year-olds).[2]

[2]Including the five participants who failed the exclusion criteria (and thus were likely not paying attention) has little influence on the results: only two provided an incorrect answer to *any* test question, both on the "singly-determined" test trial.

Discussion

Using animated physical collision displays and methods similar to those of Harris et al. (1996) and Rafetseder et al. (2013), we found a similar pattern of results to what has been reported previously: Children robustly succeeded in cases where the counterfactual question changed the outcome, but not when the outcome was over-determined. At the same time, when asked to engage in *future hypothetical* reasoning about these events, children showed no difficulty at all and were, in fact, uniformly correct.

There were some differences from previous results. Notably, there were no age effects, and rather than getting the question systematically wrong in the over-determined case, children were merely at chance. This contrasts sharply with McCormack et al.'s (2018) results who also asked for counterfactual reasoning about a physics-based event, but found much greater success at 7 and 9 years of age. However, with these methods it is impossible to say why these results were different, because we do not know why children picked the wrong answer. It is possible that they employed some alternative strategy to answer the question, or it is possible that they simulated different counterfactual alternatives than we, as adults, expected them to. Alternatively, both may have occurred. This study, like its predecessors, cannot distinguish these possibilities. Some of these possible mechanisms could have led children to a more apparently "successful" pattern in McCormack et al. (2018), even if the underlying reasoning process was the same. As a validation of the stimuli, however, Experiments 1a and 1b show that counterfactual reasoning about these physical collision events follow similar qualitative patterns as narrative studies.

Experiment 2: Multiple choice selection

Having validated this class of stimuli in Experiment 1, Experiment 2 turned to the primary goal of this project: Determining whether children answer counterfactual questions wrong because they systematically simulate different counterfactual possibilities than adults, and if so, how exactly they differ. To answer these questions we employed both a novel method, and a novel analysis strategy. We asked participants to select one of four

counterfactual alternatives for a given event using a question that did not focus exclusively on the outcome, and used a multinomial processing tree model (MPT; Riefer & Batchelder, 1988) to evaluate hypotheses about the underlying process by which specific answers were selected. Notably, we focused on age groups that have consistently struggled with counterfactual reasoning in past work (and Experiment 1): Children ages 4–6. This approach allowed us to ask whether and when children were engaging in counterfactual simulation, and if so, how exactly they might differ from adults.

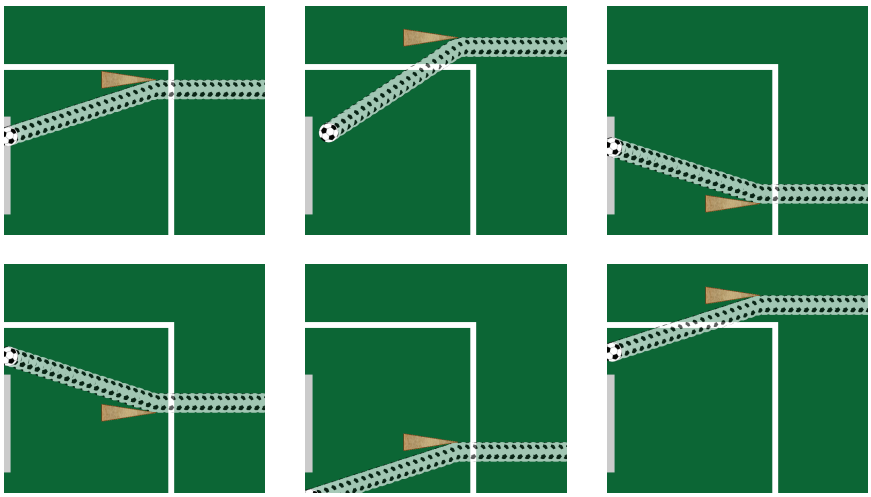
Methods

Participants. We pre-registered a planned sample size of 24 participants in each of three age groups (`https://osf.io/qn3b9`): 4-year-olds, 5-year-olds, and 6-year-olds. We therefore recruited 24 4-year-olds (15 female), 24 5-year- olds (7 female) and 24 6-year-olds (8 female). In addition, 6 4-year-olds (2 female) and 2 (female) 5-year-olds participated but were excluded because they failed to complete the study (4) or their parents interfered (3). Participants were recruited from TheChildLab.com (Sheskin & Keil, 2018).

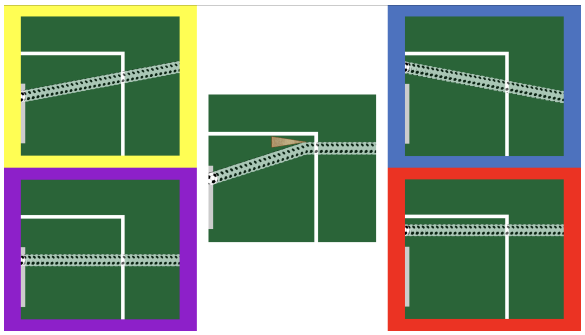
Stimuli and apparatus. Experiment 2 was approved by the Yale University IRB under protocol # 1311013027, "Cognitive and metacognitive development". Children saw a total of ten trials. Each trial first showed an animation of a physical scene. Afterwards, a grid of still images was presented. The image of what actually happened was displayed in the center of the grid, with four counterfactual possibilities presented in each corner (see Fig. 2b; full stimuli are available at `https://osf.io/5jw6y/`).

Animated events were constructed using Flash, converted to a movie format, embedded in a PowerPoint presentation, and presented over a videoconferencing system. The animations were adapted from Experiment 1. This time, there was only one ball, resembling a soccer ball, and the brick wall was replaced with a triangular wedge with a wood texture. The background was green with a white line to mimic a soccer field. The goal was turned into a grey rectangle, and there were no walls on either side of it.

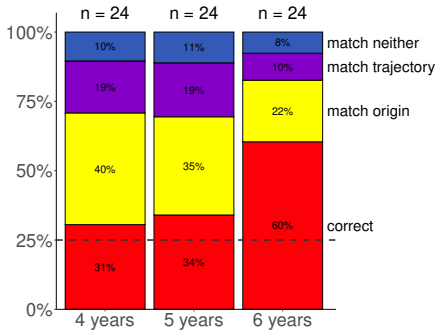
We created a total of eight test animations and two training animations. In all test



(a) Test clips



(b) Example test item



(c) Results

Figure 2. **Experiment 2**: (a) Diagrams of six test clips in Experiment 2. (b) Example test item as a child would see it. The center image shows a diagrammatic depiction of the video that the child just watched. The four images on colored backgrounds in the corners show the response options. Children were asked "If there were no block on the field, how would the ball have moved?", and answered by naming one of the colors. On this trial, red is the 'correct' response. We coded yellow as 'match origin', purple as 'match trajectory', and blue as 'match neither'. (c) Proportion of responses separated by age group. The dashed line at 25% indicates chance responding. Colors indicate the type of response, according to whether the answer preserved the origin of the ball's motion from the actual event, its trajectory, both (the 'correct' option), or neither. The colors map onto the response options show in the example test item.

animations, the ball entered the stage from the right side and moved in a perfectly horizontal trajectory. In six of the test animations, the ball deflected off of the wedge, which did (4 animations) or did not (2) change whether it went into the goal (see Fig. 2a). In two other animations (not included in analyses, see below), the ball did not interact with the wedge,

and simply moved across the field in a straight line.

Along with each test animation, we made a still image that showed the entire trajectory the ball had taken (as seen in Fig. 2a), which was visible while the child was answering the counterfactual question, thus reducing memory load. In addition, we constructed still images representing four counterfactual possibilities for each animation (Fig. 2b). In these counterfactual possibilities, the wedge was removed, and the complete trajectory of the ball was shown as in the still image of the actual event. These four possibilities were constructed in systematic ways for the six items in which the ball interacted with the wedge.

- **'Correct' (red)**: In this image, the ball starts from the same point of origin as in the actual event, and follows the same initial (horizontal) trajectory all the way to the far side of the display. This is the normatively correct option, that preserves all of the initial conditions of the actual event except for the antecedent of the counterfactual question.
- **'Match origin' (yellow)**: The ball starts from the same point of origin, but follows a diagonal, rather than horizontal, initial trajectory. The end-point of the ball's motion is in fact matched to the actual event in which it deflected off the wedge. Thus, this option preserves the origin but not the trajectory of the ball's motion in the actual event.
- **'Match trajectory' (purple)**: The ball follows a horizontal trajectory, as it does in the actual event, but starts from a different location on the right side of the display. The end-point of the ball's motion is matched to the actual event in which it deflected off the wedge. Thus, this option preserves the trajectory but not the origin of the ball's motion in the actual event.
- **'Match neither' (blue)**: The ball starts from the same location as the 'match trajectory' item, but follows an upward diagonal trajectory, ending in the same place as the 'correct' item. This option preserves neither the origin of the ball's motion nor

its initial trajectory from the actual event.

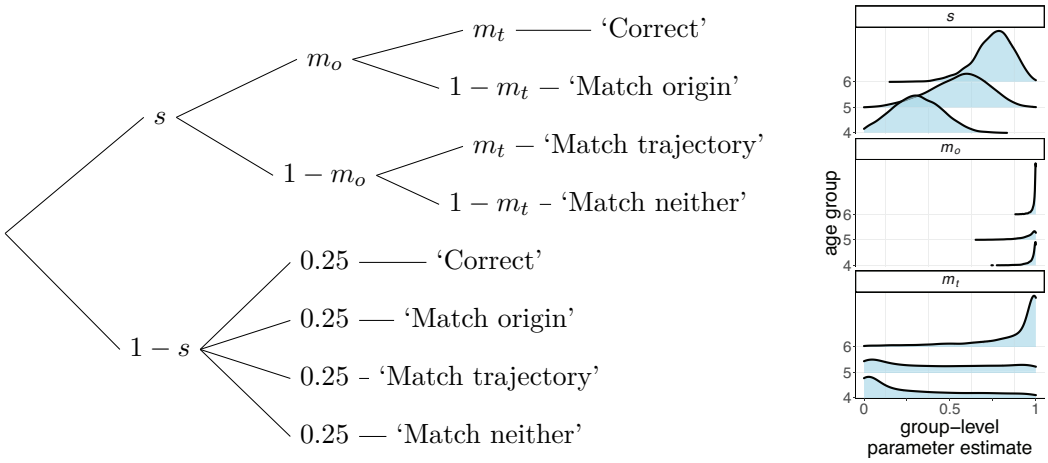
For the events in which the ball and wedge did not interact, the four images still contained two options that preserved the origin and two that preserved the trajectory, but because the ball did not deflect off the wedge in the actual event, the "match origin" and "match trajectory" images in fact showed the ball ending up in a location that was not present in the original event, while the "correct" and "match neither" images did. The model we used to analyze children's responses (described below) therefore does not apply to these trials.

In addition, there were two training animations, one in which the ball bounced off the wedge and one in which it did not interact with the wedge. In both training animations, the ball entered on a diagonal trajectory. The training task was to choose which out of four images matches the actual event. Here, there was no image of the actual event in the center of the response screen. Each of the four images showed the full trajectory of the ball with the wedge being present.

Procedure. The experimenter script can be found in the presenter notes of the PowerPoint presentations (or corresponding PDF) at `https://osf.io/5jw6y/`.

After parents gave informed consent, children were first shown the two training animations, and after each animation, they were asked to find the image that matched what they saw from the four possibilities. Children identified the image by naming the color that surrounded it (see Fig. 2b). This was primarily to familiarize children with the multiple-choice response method. For test trials, children were asked "If there were no block on the field, how would the ball have moved?".

Note that while these displays did involve a ball going into or missing a goal, and cases in which a block did or did not change that result, the question was not focused on this binary outcome. This was a deliberate choice, and a departure from past work. Rather than focusing on the outcome, we were interested in whether children created a counterfactual simulation for *the episode as a whole*, and so we asked a question that captured the entire



(a) Tree representation of the MPT model developed for Experiment 2 (b) Parameters

Figure 3. **Experiment 2 model**: a) Each path from the root node on the left to the response categories on the leaf nodes (right) represents one assumed set of cognitive steps that results in the response category. The model has three free parameters: s = probability to engage in counterfactual simulation, m_o = conditional probability to maintain the origin, and m_t = conditional probability to maintain the trajectory. The order of m_o and m_t is flexible, the model assumes these 'branches' can occur in either order. b) Posterior distribution of the group-level parameter estimates of the MPT model. The height of each distribution represents the relative evidence that the average parameter in each age group takes on this value.

episode.

The experimenter was blind to what the child was seeing at all times, and only recorded the color that the child said to identify the image. Children's responses were then transcribed by another coder who was blind to condition, and later matched to images based on the condition the child had been assigned to (see data files in repository). There were two exclusion criteria: If the child failed to finish the study for any reason, or if the parent interfered in a way that guided the child toward a specific answer on any item, in the opinion of the experimenter or coder. As both were blind to what the child was seeing, these judgments could not be influenced by knowing what option the child was selecting.

Analysis plan. Following our pre-registration, we analysed the data using a newly developed multinomial processing tree (MPT) model. MPT models are a flexible class of cognitive measurement models for categorical data that can be represented in a tree graph.

Fig. 3a shows the tree representation of our model.

An MPT model consists of a number of discrete cognitive processing steps (Riefer & Batchelder, 1988). Each model parameter represents the conditional probability of reaching a particular step or mental state. One path through the tree from the root node (on the left) to one of the observable response categories (on the right) represents one hypothesized series of steps that results in the observed behaviour. The combination of all paths that result in the same response constitute all possible ways by which a particular response can come about according to the model. For example, our model assumes that a correct response ('Correct') can either be achieved by a correct simulation (path through s, m_o, and m_t) or through guessing (path through $1 - s$ and top branch with 0.25).

The first assumption in our model is that children either engage in simulation, with probability s, or do not engage in simulation, with probability $1 - s$. In case children do not simulate we assume they randomly choose one of the four response categories (i.e., the conditional probability of choosing any one response category given that a child doesn't simulate is .25). In case children engage in simulation, we assume two further (unordered) processing steps: how likely they are to maintain the origin of the ball's movement from the actual world in their simulation (parameter m_o), and how likely they are to maintain the horizontal trajectory of the ball's movement (parameter m_t)? If children maintain both the origin and the trajectory, with probability $m_o \times m_t$, they will provide the correct response (Fig. 2b, red bottom right image). If, however, children only maintain the origin, but not the trajectory, with probability $m_o \times (1 - m_t)$, they will respond with 'Match origin' (Fig. 2b, yellow top left image). In case children do not maintain the origin, with probability $1 - m_o$, they can maintain the trajectory, with probability m_t, and respond with 'Match trajectory' (Fig. 2b, purple bottom left image), or not maintain the trajectory, with probability $1 - m_t$, and respond with 'Match neither' (Fig. 2b, blue top right image).

The model has three free parameters (s, m_o, and m_t) for three independent data points per age group (i.e., the four response categories minus one). Even though the model is saturated, it is still limited in what data patterns it can account for. For example, the model

would be unable to account for a pattern of responses where children frequently selected 'Correct' and 'Match neither', but rarely selected 'Match origin' or 'Match trajectory', because the ratio of 'Correct' versus 'Match origin' responses is determined by the same m_t parameter as the ratio of 'Match trajectory' versus 'Match neither' responses. For example, selecting 'Correct' over 'Match origin' implies $m_t > 0.5$, whereas selecting 'Match neither' over 'Match trajectory' implies $m_t < 0.5$. And both conditions cannot be true for m_t at the same time. The model only predicts a large proportion of 'Match neither' responses if m_o and m_t are both relatively low, while s is high, so a pattern of responses that was predominantly 'Correct' and 'Match neither' would be inconsistent with the model.

Consequently, our first (preregistered) hypothesis is that the assumptions of our model provide an adequate characterization of the data (i.e., that the model will fit the data). To test this hypothesis, we estimated the model using a hierarchical-Bayesian approach (Klauer, 2010) implemented via `TreeBUGS` (Heck, Arnold, & Arnold, 2017). A hierarchical-Bayesian approach allowed us to take individual differences into account even though we only have a low number of observations per child (six) by sharing information across participants.[3]

Having established that our assumption adequately describe the observed data, we can use the group-level parameter estimates per age group to distinguish between four further (pre-registered) hypotheses.

1. Children do not answer in a manner that is consistent with simulation, indicated by a small s parameter (i.e., near 0). All other hypotheses assume that s is clearly above 0 (i.e., $s >> 0$).
2. When children simulate, their simulations retain the origin of the ball's motion from the actual event, but not its trajectory, indicated in the model by $m_o >> m_t$.
3. When children simulate, their simulations retain the trajectory of the ball's motion

[3] We fit the model using four independent MCMC chains. After discarding 120,000 samples as adaptation and burn-in samples, we retained every 300th sample from an additional 300,000 samples per chain resulting in 1000 posterior samples per chain. Chain statistics (all $\hat{R} < 1.04$, all $n_{\text{eff}} > 1000$) and visual inspection indicated convergence of the model.

from the actual event but not its origin, indicated in the model by $m_t >> m_o$.

4. Children simulate in an adult-like manner and typically preserve both the origin and trajectory of the ball's motion, and typically choose the 'Correct' answer, indicated by both m_t and m_o being large (i.e., $> .5$).

To assess potential differences in parameter estimates between age groups, we calculate difference distributions between the group-level estimates. We then consider both the 80% as well as the 95% highest posterior density interval (HDI) of those difference distributions. If the 80% HDI does not contain 0, we interpret this as evidence that two parameter estimates differ across age groups.

Results

Fig. 2c shows how often children chose each of the four options for the six test items where the ball collided with the wedge. For the two cases in which the ball and wedge did not interact, the correct answer was the modal response in every age group (4-year-olds: 50%; 5-year-olds: 71%; 6-year-olds: 88%).

A visual inspection of the figure suggests a clear pattern when it comes to choosing the correct answer: It is selected at above-chance rates by age 6.[4] However, it is also evident that, of the three possible incorrect responses, all age groups preferred "match origin" over "match trajectory" and "match neither". To understand the origin of this pattern of responses, we fit our MPT model to children's responses.

Model fit was evaluated using posterior predictive p-values by comparing expected versus observed misfit (Klauer, 2010).[5] On the group-level, the model provided an adequate

[4] To test whether the actual outcome (ball going into or missing the goal) and the counterfactual outcome (block did or did not change the outcome) affected the results we analysed the accuracy (i.e., children choosing the correct response) using a binomial generalized linear mixed model (GLMM). As fixed effects we entered age group, actual outcome, and counterfactual outcome, as well as all interactions and employed both the maximal as well as a reduced random effect structure (see supplemental materials for details). Only the effect of age group, $p = .012$, as well as the age-group by counterfactual outcome interaction, $p = .009$, reached significance (all remaining $p > .1$, full details available in the supplemental analyses §5.3.2). The interaction revealed that the rate of "correct" responses increased significantly with age for the over-determined items but not for the singly-determined items, which aligns with the results of McCormack et al. (2018).

[5] H_0 for this p-value is that the observed degree of misfit is not larger than what would be expected under

account for both the mean observed category frequencies, $p = .145$, as well as the covariances among children, $p = .139$. Furthermore, on the individual-level, for none of the 72 children was the observed misfit larger than expected, smallest $p = .075$. This indicates that overall, the assumptions characterizing our model are satisfied by the data.

Fig. 3b shows the group-level posterior distributions of the model parameters. It is clear that for all age groups the s parameter – the probability to engage in counterfactual simulation – is above 0. In other words, the model estimates that all age groups engaged in simulation at least some of the time. The peak of the posterior distribution for s is lowest for the 4-year-olds ($\text{mode}_s = 0.30$, 80% HDI = [.12, .49]), and it increases with age. This increase is further supported by the HDIs of the difference distributions with the 6-year-olds. Both, the 80% HDI and the 95% HDI of the *differences* between 4-year-olds and 6-year-olds does not contain 0, indicating a particularly clear difference in this parameter between both age groups. In other words, our model indicates that 6-year-olds are more likely to engage in simulation than 4-year-olds.[6] For the 6-year-olds, the mode of s is 0.79 (80% HDI = [.63, .91]) suggesting that at this age children generally engage in counterfactual simulation. However, the 80% HDIs of the differences between the 5-year-olds and the other two age groups includes zero, indicating that the 5-year-olds are situated somewhat in the middle ($\text{mode}_s = 0.60$, 80% HDI = [.37, .79]), and their likelihood of engaging in simulation is not significantly different from that of either 4-year-olds or 6-year-olds.

For the m_o parameter – the probability to maintain the origin if children engage in simulation – the pattern is very clear. The estimated mode for all three age groups is $> .99$ with the widest 80% HDI for the 5-year-olds, [.90, 1.00]. Consequently, there are no meaningful differences between the different age groups (i.e., the 80% and 95% HDIs of all comparisons between age groups contain 0).

For the m_t parameter – the probability to maintain the trajectory if children engage

---

the model. Thus, $p < .05$ would indicate the model does not provide an adequate account to the data.

[6] We also performed a power analysis to evaluate how likely it would be to obtain such a result (i.e., a difference distribution excluding 0) given our experimental design. For the difference in s that we observed here and an 80% difference interval, the power is around 0.75. Full details are given in the supplemental analyses §5.3.3.4.

in simulation – we observe wide posterior distributions that span from 0 to 1 for all age groups. For the 4-years-olds there is a noticeable peak around 0 with considerably posterior mass at least up to 0.75. For the 5-year-olds, the peak at 0 is somewhat attenuated with considerable posterior mass over the complete range. These two results suggest that, when simulating, some 4 and 5-years-old children do not maintain the trajectory and most only do so occasionally. For the 6-year-olds, the pattern is flipped; there is a very strong peak around 1 with some posterior mass extending to 0.5. The oldest children mostly maintain the trajectory in their simulations. This differential pattern is also supported by the difference distributions. The 80% difference HDIs comparing 4-years-olds with 6-years-olds and comparing 5-years-olds with 6-years-olds do not include 0.

In terms of our pre-registered hypotheses, the different age groups also show a differential pattern. For the 4-year-olds the s parameter is comparatively low suggesting that those children only rarely engaged in any counterfactual simulations (Hypothesis 1). For both the 4 and 5-year-olds we further observe that $m_o >> m_t$ suggesting that if the younger age groups engage in simulation they are more likely to maintain the origin than the trajectory (Hypothesis 2). For the 6-years-olds, the results were largely in line with the normative solution and comparatively high value for all three parameters (i.e., all estimates near 1, with s peaking around .75; Hypothesis 4). For none of the age groups did we find evidence that in case of simulation, the trajectory is maintained, but not the origin (i.e., no support for Hypothesis 3).

The most surprising result was the extremely wide posterior distribution for m_t spanning the whole parameter range, that is, the model had a high degree of uncertainty in its estimate of m_t. The most likely culprit for this pattern were large individual differences in the m_t parameters. In line with this, the standard deviation of the *individual*-level m_t parameters was rather large; the posterior mode was around 5 [80% HDI: 3.1, 10.3] on the probit scale. Furthermore, individual estimates of m_t seemed to exhibit a bimodal pattern with peaks at 0 and 1, suggesting that some children were very likely to preserve the ball's trajectory in their counterfactuals while others almost never did (see supplemental analyses

§5.3.3).

In a final analysis step, we investigated the pattern of individual differences using a latent class approach (Klauer, 2006, see supplemental analyses §5.3.4). This approach assigns individual participants – regardless of age – into different classes which each share the same parameter values. Results indicated that three classes were required to best account for the data. The three classes differed systematically in both age composition and which hypothesis they corresponded to. Class 1 did not show substantial evidence for counterfactual simulation (Hypothesis 1: maximal uncertainty for all parameters), and encompassed around 50% of all children, these children were on average young (around 60% of 4 and 5-years-olds and 33% of 6-years-olds). Class 2 generally engaged in counterfactual simulation ($s > .5$), but only maintained the origin (Hypothesis 2: $m_o \approx 1$, $m_t \approx 0$). This class encompassed around 25% of children in the middle of our age range (25% of 4 and 5-years-olds and 17% of 6-years-olds). Children in Class 3 exhibited largely normative or adult-like behaviour (Hypothesis 4: $s \approx m_o \approx m_t \approx 1$). This class encompassed around 25% of children who were on average older (5% of 4-year-olds, 17% of 5-years-olds, and 50% of 6-years-olds).

Discussion

Children consistently selected the 'Correct' answer by 6 years of age, but even before then, they were remarkably systematic in their responses: Even younger children were much more likely to select the 'Correct' or 'Match Origin' options than the 'Match Trajectory' or 'Match Neither' options. Our model's best explanation for younger children's responses is that they did not always engage in simulation (indeed, our model suggests that many of them picked randomly), but when they did, their simulations differed from what we would expect adults to do. More concretely, younger children, like adults, seem to simulate in a way that essentially always preserves the origin of the ball's motion, but unlike adults, they allow the ball's initial trajectory to vary. Thus, we propose, young children *do* engage in counterfactual simulation, but they nonetheless answer counterfactual questions

'incorrectly' because they simulate different possibilities than an adult would.

Before discussing the implications and limitations of these findings, we must first acknowledge a salient alternative explanation for this pattern of responses. What if, rather than engaging in simulation, children simply employed a visual matching strategy? The two options children selected most often in Fig. 2b (the correct answer and the match-origin answer, red and yellow respectively), are also the two options that seem most visually similar to the event as it actually occurred. Perhaps younger children simply relied on a visual matching strategy rather than a true simulation, which would yield a very similar pattern of results (including accuracy on the items in which there is no interaction with the block). Experiment 3 was designed to test this alternative explanation.

Experiment 3: Visual matching or counterfactual simulation?

In order to determine whether the pattern of responses observed in Experiment 2 resulted from a visual matching strategy rather than from simulation, we decided to pit a true visual match against a simulation. This study, which was much more minimal in its design, took one of the items from Experiment 2 and gave 4-5-year-old children two options: A close visual match differing only in the absence of the block, or the 'correct' answer from Experiment 2 (see Fig. 4). Thus, if children use a visual matching strategy, they should overwhelmingly choose the close visual match, even though it involves a physically implausible event (a spontaneous change in trajectory with no collision). Infants are sensitive to this kind of physical violation (Kominsky & Carey, 2018; Kominsky et al., 2017), so if children engage in simulation that employs any kind of internal 'physics engine' (Ullman, Spelke, Battaglia, & Tenenbaum, 2017), they should reject this physically implausible option.

Methods

Participants. We pre-registered a planned sample of 30 4-year-olds (`https://osf.io/rd23c`), recruited from pre-schools in the greater Newark area. Our final sample consisted of 30 children age 4 years 0 months to 5 years 3 months (average age 4 years

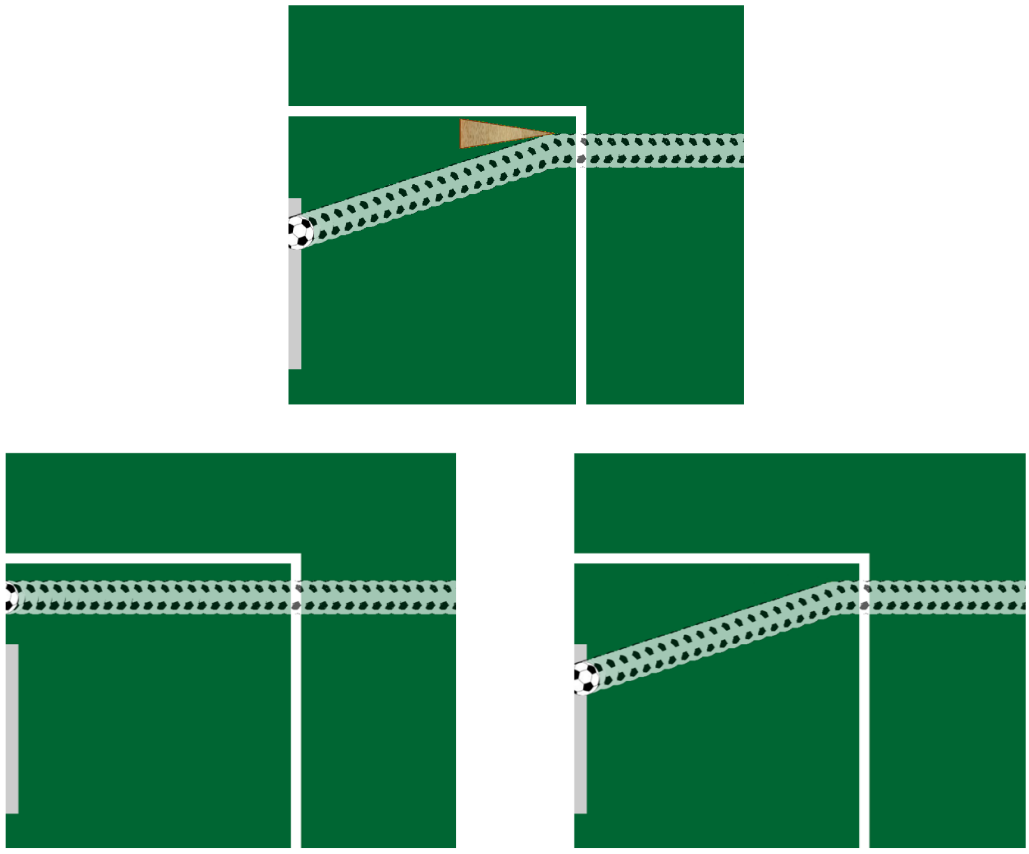


Figure 4. Stimuli for Experiment 3. Participants saw the animation of the event displayed at the top. They then saw this display and responded to the question "If there were no block on the field, how would the ball have moved?" by choosing one of the two trajectories at the bottom. Here, the "correct" answer is on the left and the "visual match" (identical to the actual event, but without the block) is on the right.

7 months; 16 male, 14 female). One additional 4-year-old participated in the study but was excluded due to failing to answer any of the questions after repeated prompting.

Stimuli and procedure. Experiment 3 was approved by the Rutgers University IRB under protocol 2020000399, "Learning, perception, and belief revision". This study was conducted on a tablet using the Qualtrics offline app (which did not exist when we ran Experiment 1). Each child participated in this experiment immediately following their participation in an unrelated study which involved neither collisions nor physics nor counterfactual questions. Participants first saw a training item in which they were told a story about a girl who dropped an ice-cream cone. They were asked whether the girl would be

happy or sad if she had not dropped the ice-cream cone. This was to familiarize children with the response method and counterfactual wording. Participants were not corrected if they answered "sad".

Participants then received the same initial instructions as in Experiment 2, without the two training items. After instructions, they saw one of the animations from Experiment 2 (the one shown at the center of Fig. 2b), and were presented with a screen that showed the event as it actually occurred with two options (presented in Fig. 4): The 'correct' answer (same as the 'correct' answer in Experiment 2), and a 'visual match' option, which showed the ball following the exact same trajectory as it had in the actual event, but without the block on the field (so the ball appeared to spontaneously changes direction). Children were asked the same question as in Experiment 2 ("If there were no block on the field, how would the ball have moved?").

For both the training and test item, the side of each choice was randomized. To ensure that experimenter bias could not influence children's responses, the experimenter turned the tablet so they could not see the screen before presenting the two options, so the experimenter was blind to which response option was presented on which side. Children indicated their choice by pointing to or touching one of the two options on the tablet screen.

Results

Precisely 20 of our 30 participants selected the 'correct' option on the test trial, while the other 10 selected the 'visual match' item. In an exact binomial test this is not significantly different from chance responding (50%), $p = .099$, but the goal of this experiment was not to compare performance against chance. Rather, it was to compare our simulation-based explanation from Experiment 2 against a direct visual matching strategy.

To test whether our results were consistent with the estimated rate of engaging in simulation in Experiment 2, we conducted another MPT analysis with only the s parameter (as there were only two response options, we could only estimate one parameter). The model for this pattern of responses is very simple: We assume that if children are engaging

in simulation, they will select the 'correct' answer. If they are not engaging in simulation, we assume they will choose randomly. Thus the probability of picking the correct answer is $s + (1 - s) \times .5$ and the probability of picking the incorrect answer is simply $(1 - s) \times .5$. If our parameter estimate is comparable to what we see in 4-year-olds in Experiment 2, it suggests that the presence of a direct visual match did not change how children responded.

Indeed, a Bayesian model estimate in this experiment ($\text{Mode}_s = 0.34$, [80% HDI: .15, .54]) was highly similar to the estimated mode for the 4-year-olds in Experiment 2 ($\text{Mode}_s = 0.27$, [80% HDI: .12, .39]), which fits our account of these results: 4-year-olds do engage in simulations to answer these questions roughly a third of the time, and when they do not, they pick randomly.[7]

Discussion

It is unlikely that the systematic nature of children's wrong answers in Experiment 2 are the result of a simple visual matching strategy. Even when presented with a perfect visual match, a majority of children chose the simulation-consistent response, and adding a perfect visual match did not alter the estimated rate of engaging in simulation in our MPT model for this age group compared to 4-year-olds in Experiment 2.

General discussion

Children often answer counterfactual questions incorrectly, especially prior to age 6. We have provided the first evidence that, at least some of the time, these 'incorrect' answers are consistent with the possibility that children engage in counterfactual simulation, but simulate different possibilities than we would as adults. Experiments 1a and 1b extended children's failures in counterfactual reasoning tasks and successes in hypothetical reasoning tasks to the domain of simple collision interactions. In counterfactual reasoning, children are accurate when the outcome would have changed in the counterfactual situation, but

[7] Because we only have one trial per child, we opted for the Bayesian model on the aggregated data in this case (instead of a hierarchical-Bayesian model over individual data). This assumes that individual variability is consistent with a multinomial distribution.

struggle when the outcome was over-determined. However, they nonetheless are perfectly accurate when making predictions for both types of cases. Experiment 2 showed that, when allowed to pick specific counterfactual possibilities, children's "wrong" answers seem to result from a process of simulation that selectively preserves the origin of an object's motion, but not its initial trajectory. Experiment 3 provided evidence against the alternative account that children's responses could be explained by a simple visual matching strategy. The experiment pitted a precise visual match against a simulation-consistent response and showed that 4-year-olds responses still followed the predictions of our model, and were consistent with their responses being driven by counterfactual simulation rather than merely relying on visual similarity.

The development of counterfactual simulation

Previous work on the development of counterfactual reasoning has focused on when the ability to engage in counterfactual simulation emerges (Beck & Riggs, 2014; Rafetseder & Perner, 2018; Rafetseder et al., 2013). Our results concur that there are developmental changes in the *use* of simulation, but also show that we must examine changes within the *process* of counterfactual simulation as well. The use of relatively coarse outcome-focused measures in past work makes it difficult to isolate these different developmental contributions, which may be part of the reason for the large variance in when children succeed at counterfactual reasoning between different studies. Because past work did not account or look for the possibility that children were simulating in a non-adult matter, in some cases such simulations may have led children to the 'correct' answer, while in others it may have led them astray. Using dynamic collision events as stimuli, we were able to identify a clear developmental change in counterfactual simulations between ages 4-5 and 6: Younger children are likely to preserve the point of origin of an object's motion when they conduct a counterfactual simulation, but allow the object's initial trajectory to vary, whereas older children are likely to preserve both. This raises three key questions: First, why do children allow the initial trajectory to vary? Second, why do they preserve the point

of origin in particular? Third, what exactly is changing between how children and adults simulate counterfactuals?

Recent work has suggested that, in the context of causal reasoning, children have a wider and "flatter" hypothesis space than adults (i.e., priors across all hypotheses are similar), in which they conduct a "higher-temperature" (i.e., broader) search (Gopnik et al., 2017). Counterfactual reasoning in adults has been modeled as a sampling process over a distribution of possible worlds (Gerstenberg et al., submitted, 2017; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Icard et al., 2017; Kominsky & Phillips, 2019), and so it is straightforward to extend Gopnik et al.'s (2017) proposal to counterfactual reasoning: The space of counterfactual possibilities that children sample from is flatter, and their sampling process is broader. However, while this view offers one explanation for why children's responses vary more than adults in general, it does not explain why they vary in this particular way. There is nothing in this view that predicts that certain responses are more likely to be selected than others.

In general, the reason children might preserve the origin of the ball's motion is that they treat the origin as a 'background condition' rather than a 'mutable' feature of the event (Byrne, 2016). As a classic example of the difference, consider a forest fire: What caused the fire to start, a lit match, or the presence of oxygen in the air? The presence of oxygen is not considered a 'mutable' part of this event (McGill & Tenbrunsel, 2000). Instead, it is one of the assumptions we make about the state of the world. In one sense, to consider a possibility in which this condition is changed is to abandon the premises of what actually happened and construct a different event altogether. As such, children might view the point of origin as a precondition for the event, but the trajectory as something that could be changed. Indeed, recent work has suggested that whole types of events that adults seldom treat as mutable are considered to be mutable by younger children, for example, the weather (Nyhout & Ganea, 2020).

But *why* treat the origin in particular, and not the trajectory, as a background condition? One possibility is that children view parts of the episode that are further back in

time as less mutable. This would fit with work that has argued that adults' causal and counterfactual judgments tend to focus on the most recent event in a causal chain that could change the outcome (Byrne, 2005). In these events, we would regard the most recent antecedent of the outcome to be the collision with the block, which is altered in every option children are given. However, children may go back one step further and change the trajectory as well. However, some studies have found that children tend to change earlier rather than later events in a scenario (Rafetseder, Cristi-Vargas, & Perner, 2010). So it is possible that temporal order is not the critical feature.

Notably, children aren't just arbitrarily picking one feature of the original event to preserve: they selected the option that preserved the trajectory but *not* the origin ('Match trajectory') almost as infrequently as they selected the option that preserved neither. An alternative possibility is that children are inferring an unseen source of the ball's motion that can change the trajectory easily, such as a person kicking the ball in a different direction. This could be viewed as children forming a different causal model of the event than an adult would (Nyhout & Ganea, 2019), or even a form of pretense, which is a kind of simulation (Buchsbaum et al., 2012). Pretense is typically distinguished from counterfactual reasoning by being less anchored to reality: a pretend possibility does not need to connect to events as they actually occurred, and allows for the creation of entirely new factors that were absent from the real world. However, it is possible that children engage in a sort of hybrid simulation that has some features of pretense and some features of counterfactual reasoning. At this point, such an account is mostly speculation. While we cannot yet be sure of the reason, we can say that children find the trajectory of the ball's motion to be mutable in a way that its point of origin is not.

Accounts of developmental change

There are other accounts of how children's counterfactual reasoning differs from that of adults' that fit with our results to varying degrees. One explanation is that unlike adults, children have difficulty maintaining multiple possibilities in mind at the same time (Carey et

al., 2020), that is, they cannot consider both reality and the counterfactual possibility at the same time (Beck et al., 2006). We believe there is some merit to this proposal, but it is not a complete explanation. Experiment 2 specifically attempted to remove this challenge by presenting children with both reality and multiple counterfactual possibilities at the same time, essentially removing the cognitive load of maintaining two possibilities by putting the possibilities in front of them. While this likely improved their performance relative to a task where they had to maintain the episode as it actually occurred in memory alone, it did not lead to consistent success. Another explanation is that children have difficulty inhibiting reality in order to consider the counterfactual (Beck & Riggs, 2014). Our findings (particularly those of Experiment 1a and Experiment 3) do not suggest a reality bias, but once again the way in which we presented our task may have made this less of a challenge, as children were able to see both reality and the alternatives at the same time, permitting them to conduct easy contrasts between the two. Thus we do not reject any of these three of these explanations, but suggest that none of them can fully explain what changes between children and adults.

Another family of developmental explanations focuses on a shift in the type of reasoning employed, the move from basic conditional reasoning (BCR) to true counterfactual reasoning (Leahy et al., 2014; Rafetseder et al., 2013). This is partly compatible with our results, but in general children in this study did not respond the way we believe a BCR account would predict. The key feature of BCR is that the actual state of anything mutable is ignored, and is re-generated either based on the antecedent of a counterfactual question or from logical reasoning from immutable aspects of the scenario. That was an assumption our model considered: If they did not engage in simulation, we expected them to allow everything in the scenario to vary, that is, choose randomly among all four options. While we have argued that, as it happens, children are treating the origin as less mutable than the trajectory, we provided them with options that change the origin as well, and nothing in the counterfactual antecedent specifies that the origin cannot be changed. In other words, because they had the option of changing the origin, they should have been willing to do

so, under a BCR account (or at least, we see no reason they would not). However, our model also suggested that children did not always use simulation to answer these questions, and indeed, they did select options that changed the origin some proportion of the time. Thus, while we have argued that one source of developmental change is a change in the process of counterfactual simulation itself, our findings suggest that the frequency of using counterfactual simulation changes over development as well.

Limitations and open questions

These experiments provide an initial test of the proposal that children engage in counterfactual reasoning but do so in a non-adult-like manner. Like all initial tests of novel proposals, there are limitations that must be addressed in future work. The first, and most obvious, is that we focused exclusively on a domain of simple physical interactions. Experiments 1a-b showed that it is possible to replicate the pattern of results in this physical domain that previous studies have found in scenarios involving agents. However, it does not follow that the same pattern of selectively variable simulation that we observed here would also appear in those other domains, that is, the preservation of some 'origin' equivalent while allowing a 'trajectory' equivalent to vary, in cases where such concepts may not apply. Indeed, one other study with narrative stimuli has used a similar multiple-choice paradigm to the one we employed in Experiment 2 (Rafetseder & Perner, 2018). However, the options they provided focused exclusively on the outcome and did not vary along the same dimensions (i.e., their choices were not generated by manipulating 'origin' and 'trajectory' or analogous features), and their analyses focused primarily on whether children chose the correct answer, and the relationship between their answer and measures of false belief. They did find success at a similar age (slightly younger, in fact, with substantial success among older 5-year-olds), but it is otherwise difficult to draw a direct comparison between our results and theirs. That said, one challenge to conducting a more directly comparable study outside the domain of our physics-based stimuli is that the relevant counterfactual possibilities are well-defined for simple physical events, while the scope of counterfactual pos-

sibilities that children might generate in narrative cases involving intentional agents could be much less constrained, and therefore harder to capture systematically in a forced-choice task.

In addition to using a different response paradigm, Experiments 2 and 3 used a different question than has been studied in past work. Experiment 1, and much previous work, asked questions that focused on the outcome, "would the ball have gone into the goal?" or "would the floor be dirty?". We asked "How would the ball have moved?", which was intended to prompt children to consider the whole episode rather than just whether the ball went into the goal or not. There are several potential ramifications to using this question instead of an outcome-focused question. For one, it meant that the distinction between "over-determined" and "singly-determined" events was not the focus of Experiment 2, even though we did incorporate this feature into our stimuli (though a supplementary analysis did find some developmental differences in the rate of "correct" answers, see footnote 4). However, the primary influence of the question may have been that it presented a "how" counterfactual rather than a "whether" counterfactual (Gerstenberg et al., 2015). There is a possibility that children use different strategies to answer these questions, and that they may in fact have been more inclined to engage in simulation because it was a "how" counterfactual than they would have otherwise. On the other hand, given that we found a developmental pattern that aligns with several studies using "whether" counterfactuals (McCormack et al., 2018; Rafetseder & Perner, 2018, e.g.,), we feel it is more likely that we have captured a general developmental trajectory for counterfactual reasoning which relies both on a growing ability to simulate and a growing ability to simulate in an adult-like manner.

Another limitation, but also a potential strength, is that we provided the specific counterfactual alternatives that children had to choose between. Our interpretation is certainly limited by which possibilities children had available to them. One might argue, for example, that children were not concerned with the 'origin' and 'trajectory' as much as where the ball started and where it ended up. However, the "Match trajectory" and "Match origin"

options in Experiment 2 both matched the outcome of the ball's motion, and yet children rarely chose the "Match trajectory" response. Even so, one could reasonably wonder if they would have selected the "Match origin" option as much if the ball also ended up in a different location altogether. If changing the "Match origin" option in this way led to a different pattern of responses, it would suggest a much more complex relationship between the origin, trajectory, and outcome in how children consider counterfactual possibilities. There were also a number of other parameters of this episode that we could have manipulated and did not (e.g., changing the angle of the block rather than removing it altogether) or that could be challenging to render with the stimuli we used (e.g., changing the ball's speed would be difficult to capture in the still images we used). Put simply, there are many counterfactual possibilities that children could have considered that we would have been unable to capture with these methods.

In our experiment, we infer that children engaged in counterfactual simulation from the concrete response options that they chose. Future work should study the process of mental simulation more directly. For example, children could draw the trajectory of the ball's movement on a blank field, rather than be presented with predetermined options. In fact, we piloted such a procedure, but ultimately ran into the twin problems that 1) children often got distracted by the act of drawing itself and 2) while young children definitely *enjoy* drawing on a tablet, a motor skills check item revealed that they were so imprecise that anything they drew would be very difficult to interpret. For example, based on a children's drawn path it would be difficult to tell whether it should be characterized as "Correct" versus "Match origin". These two responses would be within the margin of error for younger children's motor skills, based on a tracing task we included in our pilot experiment. However, future work may be able to circumvent this problem by using either much larger displays where the observed motor noise does not swamp responses, or by using eye-tracking measures. An eye-tracking study building on the current methods might also allow us to investigate the time-course of children's simulations: do children, like adults, simulate counterfactual possibilities while watching the events unfold (Gerstenberg et al.,

2017)? Do children's eye-movements suggest that they are more likely to change the trajectory? While the current data do not answer these questions, this study lays a clear path for future work along these lines.

The strength of our multiple choice response paradigm is that it allows us to systematically vary different features of the episode and pit them against each other. In this experiment, we focused on two features of the event in creating these options (the point of origin and initial trajectory). Future work could focus on other dimensions of these displays, or of whatever sort of displays are used. This does constrain the conclusions we can draw — for example, it is not the case that every failure of counterfactual reasoning can be attributed to a tendency to preserve origin but not trajectory — but it does give us a more precise understanding of the process by which these counterfactuals are generated. Many more studies that look at different dimensions in greater detail will be required to fully understand this process, even for simple physical events like the ones studied here. The goal of this paper was not to provide such a comprehensive understanding, but rather to allow these questions to be asked in the first place. This work provides initial evidence that children do (sometimes) engage in counterfactual reasoning even when they answer counterfactual questions incorrectly, because they simulate different counterfactual possibilities than adults do.

Finally, there are unresolved questions, which we did not attempt to address, about how children answer when they do *not* engage in simulation. While we provide evidence suggesting that children are capable of counterfactual simulation (thus contradicting Piaget), we nonetheless believe that they may sometimes use alternative strategies to answer questions about counterfactuals. We cannot explain *all* of children's struggles in past studies as the result of systematic differences in what they simulate. We are offering a partial explanation, but further work is needed to understand how children answer these questions without using simulation. While existing theories like "basic conditional reasoning" offer one possible explanation for past results, there are others that remain untested. For example, to return to a case from the introduction, merely asking the question "What if Carol

had taken her shoes off?" may be enough to make some children think the outcome must be different than what they observed (Bonawitz, Shafto, Yu, Gonzalez, & Bridgers, 2020). A complete understanding of the development of counterfactual reasoning will require both further investigations of children's counterfactual simulation as well as the other strategies children might employ in answering these questions.

Conclusion

Over the past decade it has become increasingly clear that children are able to engage in sophisticated counterfactual reasoning between four and six years of age. Most accounts of children's failures have focused on the idea that, prior to whatever age of success is found, they simply do not engage in counterfactual simulation. Here, we find that children's responses are systematic and consistent with counterfactual simulation, but that there may be developmental differences in what counterfactual possibilities children consider. Understanding why and how children's counterfactual simulations differ from those of adults will not only help us understand the development of children's more general reasoning abilities, it will help us understand the process of counterfactual simulation itself.

Acknowledgements

We would like to thank Elizabeth Bonawitz and the Computational Cognitive Development Lab for providing feedback on the project as a whole and data collection opportunities for Experiment 3.

This research was supported by NSF grant DRL 1561143 awarded to Frank C. Keil, NIH NRSA F32HD089595 to Jonathan F. Kominsky, and Swiss National Science Foundation (SNSF) grant 100014_179121 to Henrik Singmann.

References

Atance, C. M., & O'Neill, D. K. (2005). The emergence of episodic future thinking in humans. *Learning and Motivation*, *36*(2), 126–144.

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

Beck, S. R., & Riggs, K. J. (2014). Developing thoughts about what might have been. *Child Development Perspectives*, *8*(3), 175–179.

Beck, S. R., Riggs, K. J., & Gorniak, S. L. (2009). Relating developments in children's counterfactual thinking and executive functions. *Thinking & Reasoning*, *15*(4), 337–354.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, *77*(2), 413–426.

Beck, S. R., Weisberg, D. P., Burns, P., & Riggs, K. J. (2014, August). Conditional Reasoning and Emotional Experience: A Review of the Development of Counterfactual Thinking. *Studia Logica*, *102*(4), 673–689. doi: 10.1007/s11225-013-9508-1

Bonawitz, E., Shafto, P., Yu, Y., Gonzalez, A., & Bridgers, S. (2020). Children change their answers in response to neutral follow-up questions by a knowledgeable asker. *Cognitive Science*, *44*(1).

Buchsbaum, D., Bridgers, S., Skolnick Weisberg, D., & Gopnik, A. (2012). The power of possibility: causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B*, *367*(1599), 2202–2212.

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

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality.* MIT Press.

Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could it be so? the cognitive science of possibility. *Trends in Cognitive Sciences*, *24*(1), 3–4.

Carlson, S. M., White, R. E., & Davis-Unger, A. (2014). Evidence for a relation between executive function and pretense representation in preschool children. *Cognitive*

Development, *29*, 1-16.

Ferrell, J. M., Guttentag, R. E., & Gredlein, J. M. (2009). Children's understanding of counterfactual emotions: age differences, individual differences, and the effects of counterfactual-information salience. *British Journal of Developmental Psychology*, *27*, 569–585.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (submitted). A counterfactual simulation model of causal judgment.
(preprint on OSF at `https://psyarxiv.com/7zj94/`)

Gerstenberg, T., & Icard, T. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*.

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

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

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

Heck, D. W., Arnold, N. R., & Arnold, D. (2017). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, 1–21. doi: 10.3758/s13428-017-0869-7

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019, September). A

counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164. doi: 10.1016/j.cognition.2019.05.006

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. Retrieved from `https://doi.org/10.1016%2Fj.cognition.2017.01.010` doi: 10.1016/j.cognition.2017.01.010

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence.* New York, NY: Basic Books.

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.

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*(1), 7–31. doi: 10.1007/s11336-004-1188-3

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. doi: 10.1007/s11336-009-9141-0

Kominsky, J. F., & Carey, S. (2018, July). Early-developing causal perception is sensitive to multiple physical constraints. In T. Rogers, M. Rau, J. Zhu, & C. Kalish (Eds.), *Cogsci 2018* (pp. 622–627).

Kominsky, J. F., & Phillips, J. (2019, Oct). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, *43*(11). Retrieved from `http://dx.doi.org/10.1111/cogs.12792` doi: 10.1111/cogs.12792

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kominsky, J. F., Strickland, B., Wertz, A., Elsner, C., Wynn, K., & Keil, F. (2017). Categories and constraints in causal perception. *Psychological Science*, *28*(11), 1649–1662.

Leahy, B., Rafetseder, E., & Perner, J. (2014). Basic conditional reasoning: How children mimic counterfactual reasoning. *Studia Logica*, *102*(4), 793–810.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

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

Mahr, J. B. (2020). The dimensions of episodic simulation. *Cognition*, *196*, 104085.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018, jan). 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, August). 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

McEleney, A., & Byrne, R. M. J. (2006, May). Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning*, *12*(2), 235-255. Retrieved from `http://dx.doi.org/10.1080/13546780500317897` doi: 10.1080/13546780500317897

McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of Personality and Social Psychology*, *79*(5), 677-689. Retrieved from `http://dx.doi.org/10.1037/0022-3514.79.5.677` doi: 10.1037/0022-3514.79.5.677

Nyhout, A., & Ganea, P. A. (2019, February). 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.

Nyhout, A., Henke, L., & Ganea, P. A. (2019, August). Children's counterfactual reasoning about causally overdetermined events. *Child Development*, *90*, 610-622. 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.

Payir, A., & Guttentag, R. (2019). Counterfactual thinking and age differences in judgments

of regret and blame. *Journal of Experimental Child Psychology*, *183*, 261–275.

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

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

Qualtrics. (2005). *Qualtrics*. Provo, UT.

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. (2018, April). Belief and counterfactuality: A teleological theory of belief attribution. *Zeitschrift für Psychologie*, *226*(2), 110-121. doi: 10.1027/2151-2604/a000327

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

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–148.

Saxe, R., & Carey, S. (2006, sep). The perception of causality in infancy. *Acta Psychologica*, *123*(1-2), 144–165. Retrieved from `https://doi.org/10.1016%2Fj.actpsy.2006.05.005` doi: 10.1016/j.actpsy.2006.05.005

Sheskin, M., & Keil, F. (2018). Thechildlab. com a video chat platform for developmental research. Retrieved from `psyarxiv.com/rn7w5`

Sloman, S. A., & Lagnado, D. A. (2005). Do we 'do'? *Cognitive Science*, *29*(1), 5–39.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, sep). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. Retrieved from `https://doi.org/10.1016%2Fj.tics.2017.05.012` doi: 10.1016/j.tics.2017.05.012