Reasoning about similar causal structures among mechanical systems

Alexandra Rett (arett@ucsd.edu)[1], Micah Goldwater (micah.goldwater@sydney.edu.au)[2], & Caren M. Walker (carenwalker@ucsd.edu)[1]

[1] Department of Psychology, UC San Diego, La Jolla, CA 92092

[2] School of Psychology, The University of Sydney, Sydney, Australia

Abstract

Across two experiments ($N = 256$), we test children's ability to recognize similar causal structures among mechanical systems. In Experiment 1, 4- to 7-year-olds were shown unique sets of three machine types (a causal chain, a common effect, and a common cause) and asked to judge which machines were most similar. We find that 6- to 7-year-olds, but not 4 to 5-year-olds, spontaneously match machines that share the same causal structure. However, all children relied primarily on timing cues when making similarity judgments. In Experiment 2, we control for timing cues, instead asking children to discriminate causal structure by observing an intervention on each machine. We find that, in the absence of perceptual cues, only 8- and 9-year-olds successfully matched machines based on structural similarity. We discuss potential explanations for these findings and consider ways to support recognition of common causal structure in the learning environment.

Keywords: abstract reasoning; similarity; causal reasoning; perception; causal system categories

Introduction

Reasoning About Causal System Categories

Suppose we observe that every time we start a car, the engine runs and the headlights turn on. What is the causal structure of the relationships between these variables? One possibility is that starting the car is a common cause of the engine running and the headlights turning on (starting the car activates both systems independently). An alternative is that starting the car causes the engine to run, and the engine running powers the electrical system, which then causes the headlights to turn on—a causal chain.

A deep understanding of any phenomenon requires understanding the abstract causal structure that underlies it. Indeed, national science standards in the U.S. highlight the importance of making abstract connections across domains in primary school (NRC, 2012). Here, we examine the emergence of this ability to recognize instances of the same causal structure across superficially distinct events. That is, we explore the development of children's sensitivity to *causal system categories*, or abstract patterns of causation (Rottman et al., 2012). While there has been a considerable amount of research on young children's reasoning about individual causal systems (e.g., causal chain, common cause, Frosch et al., 2012; Kushnir et al., 2003; Lapidow & Walker, 2020; McCormack et al., 2016; Nussenbaum et al., 2020; Schulz et al., 2007), research on the recognition of causal system categories has primarily been conducted with adults.

Abstract causal understanding in adults is often measured based on how they choose to categorize events or problems. When asked to categorize events, novice adult learners typically categorize based on superficial properties (e.g., topic or domain), while more advanced adult learners rely primarily on underlying causal structure (e.g., causal chain or positive feedback loop) (Rottman et al., 2012; see also Chi et al., 1981; Galloway et al., 2018; Stains & Talanquer, 2008). This shift from relying on surface similarities to privileging underlying structures is linked to increased expertise, allowing learners to identify transferable causal patterns across superficially distinct events (Rottman et al., 2012; Goldwater & Gentner, 2015).

Children's Abstract Causal Reasoning

Given the difficulty of prioritizing abstract causal information over superficial similarities, we investigate whether children use causal structure as a metric for similarity while controlling for perceptual cues (e.g., color, shape, function). To our knowledge, only one prior study has examined the development of this ability in children. Specifically, Rett and colleagues (2024) presented 5-7 year olds with simple, three-variable narratives in which the story events unfold according to a causal chain or a common effect structure. When children were asked to make judgments about which stories were the most similar, 6- and 7-year-olds, but not 5-year-olds, reliably used causal structure as a metric for similarity.

While this prior work establishes that children as young as 6 can categorize events by common effect and causal chain structures, key questions remain. Specifically, in order to recognize shared causal structure across events, children must first infer the causal structure of each individual event. Learners use various cues to infer causal structure, but some are less relevant among narratives. For example, prior work has found that both children and adults often privilege temporal information when differentiating between a causal chain and a common cause structure (McCormack et al., 2009; Frosch et al., 2012; Lagnado & Sloman, 2004; 2006). Consider again the problem of inferring the causal relationships among the following events: turning on a car, the engine running, and the headlights turning on. If the engine running *precedes* the headlights turning on, one might assume the engine *caused* the lights to turn on. However, if the lights still turn on after removing the engine, this suggests that starting the car is the common cause of both events. This intervention—removing the engine—clarifies the system's true causal structure, even when it conflicts with lower-level temporal information. Thus, children's ability to infer an event's causal structure varies by the cues available.

Adults can effortlessly integrate information about timing with evidence from informative actions to infer causal structure, even without knowing the underlying mechanism. Children, on the other hand, sometimes struggle to do so, finding it easier to rely on temporal cues alone. For example,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

McCormack and colleagues (2015; 2016) found that 7- and 8-year-olds were unable to use information from interventions to accurately judge whether a three-variable causal system was a common cause or causal chain, arguing that the challenge may stem from children's difficulties integrating evidence across separate observations. Therefore, although even preschool-aged children are capable of inferring causal structure based on intervention data (Schulz et al., 2007; Lapidow & Walker, 2020), it is possible that they would preferentially rely on perceptual heuristics.

The Current Study

Here we explore whether asking children to infer shared causal structure across mechanical systems reveals the same developmental trajectory as previous work (Rett et al., 2024). The use of mechanical systems (rather than narratives) allows us to examine whether and how children integrate perceptual cues to causality (e.g., timing), with intervention information when reasoning about causal system categories.

Based on prior work, there are two competing hypotheses about children's performance. First, it is possible that children will struggle to use causal structure as a metric for similarity when comparing mechanical systems until later in development. As in earlier studies examining children's ability to discriminate between individual causal systems, this challenge could stem from difficulty integrating different cues to causal structure. Issues may also arise as a result of children's tendency to group or categorize objects based on their shared functional properties (see Gelman & Meyer, 2011, for a review), which may interfere with attention to common causal structure. Alternatively, recognizing shared causal structure may be easier in the context of mechanical systems, compared to narratives. After all, prior research on children's causal reasoning often relies on simple machines because they reduce cognitive demands and the need for prior knowledge, and highlight the relations between causes and effects (e.g., Schulz et al., 2007). In contrast, reasoning about causal structures in narratives is more challenging, requiring that children recall abstract patterns across stories, while ignoring surface details—something they often struggle to do (Williams et al., 2002; Walker & Lombrozo, 2017).

In the current experiments, children view sets of three mechanical systems and are asked to match the two they find most similar. While each trial includes only two machine types, all machines function according to either a common cause (A←B→C), a common effect (conjunctive) structure (A→C←B), or a causal chain (A→B→C). In Experiment 1, three-variable machines allow matching based on temporal cues and the number of simultaneous or sequential events. In Experiment 2, we control for temporal cues, using interventions to convey causal structure. Across experiments, we never prompt children to attend to causal structure. Instead, we examine whether they spontaneously recognize shared structure across instances. Finally, rather than asking whether children *prioritize* similarity of abstract structure over competing surface cues, we examine whether they treat causal structure as a reasonable metric for similarity in the absence of competing low-level features.

Experiment 1

Participants

A total of 128 4–7-year-olds participated, with 32 children in each age group (4-year-olds: M = 4.48 years, SD = .30; 5-year-olds: M = 5.48 years, SD = .28; 6-year-olds: M = 6.54 years, SD = .29; 7-year-olds: M = 7.41 years, SD = .34). This sample size was preregistered and based on a power analysis using data from an initial study using a similar paradigm. Based on this power analysis, 128 participants are necessary to achieve a power of .8 for the estimated effect size (odds ratio = 1.53) of the interaction between age and the type of causal structure comparison. An additional 15 children were tested but excluded due to parental or sibling interference (n = 2), technical difficulties (n = 1), excessive distraction (n = 1), or failure to complete the study (n = 11).

The study was conducted online, with data collected from the Children Helping Science (CHS) participant database (formerly Lookit), an online platform for developmental science (Scott & Schulz, 2017). CHS was used for participant management, study creation, and data collection.

Materials

A text file of the JSON for this experiment and a recording of the participant perspective of the study can be found at https://osf.io/sqt7m/?view_only=298bc0dfbcd144dbade95232ab4d08b2. The study design was implemented using Children Helping Science (Scott & Schulz, 2017).

We created six sets of machines, with three machines of each structure. Each individual machine had three lights on the top and worked according to one of three causal structure types: a causal chain, a common cause, or a common effect. The six sets of three machines were presented in six separate trials. All three machines within the same trial had the same shaped lights (e.g. star-shaped), and each individual machine was one of three triadic colors that were equidistant on the color wheel (e.g., blue, yellow and red). Across the six trials, none of the light shapes were repeated.

Each machine was shown to children in the form of a brief, realistic 3D video, created using Blender. The timing between each machine's activation was also controlled, such that each light turned on 1 second after the previous light or action. Each machine also has a small "on" button on the side of the machine that children saw a cartoon hand press to start the machine in each video.

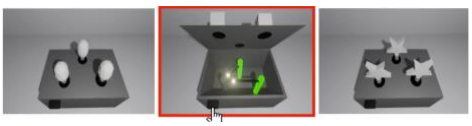

Figure 1. Exp. 1 machine demonstration. Image shows a still frame of the common effect internal mechanism.

Procedure

Warm-Up & Stimuli Preview Children participated remotely with their parent or legal guardian on CHS. No

experimenter was present during the study. Children were told they would play a matching game. To teach children how to make a selection on the screen, they first completed a practice trial. During the practice trial, they were shown an image of a dolphin, a bird, and a sailboat. They were given audio instructions to indicate whether the bird or the sailboat was most like the dolphin by clicking on the corresponding image on the screen.

After the practice trial, children were introduced to the machines. To avoid highlighting specific attributes, we introduced a variety of features of the machines. However, similarity judgments could only be based on the machines' causal structure, as there were no shared color or shapes. For each feature, children were told that machines could take on a range of different values, such as lights of different shapes (round, cube-shaped, star-shaped, etc.) and different colors (red, orange, blue, etc.), and that the machines worked in different ways on the inside. To demonstrate how the machines work, children were shown the lights activating when the machine was closed, followed by a demonstration of the machine's internal mechanism (see Figure 1). A verbal description of how the machine worked accompanied each video. For instance, for a machine with a common effect structure, children were told, "For some machines, when you push the button, two lights turn on first, which both make one light turn on. Let's see how this one works on the inside!"

Similarity Judgments After seeing the range of features that varied across machines, the game began. Children were not told which features to focus on. In each trial, three machines were shown on the screen, one at a time. Two of the machines shared the same causal structure (e.g., common effect), while the third had a different causal structure (e.g., common cause). The first two machines appeared at the top of the screen, and children were asked to decide which one was most like the third machine, which appeared at the bottom. Audio cues directed children's attention to different parts of the screen during the video demonstrations (e.g., "Look at this machine!"). Each machine's demonstration was shown twice. Afterward, children clicked on the image of the machine they thought was "most like" the machine on the bottom of the screen. They received neutral feedback ("Thank you!") and clicked to move to the next page.

The experiment consisted of six trials in which children selected which machines were most similar. Between trials, a progress page informed them how many trials remained. In each trial, two machines shared the same causal structure, while the third had a different structure. Across the six trials, there were two comparisons each of common effect vs. causal chain machines, common effect vs. common cause machines, and common cause vs. causal chain machines.

Counterfactual Questions After completing all six similarity judgments, children were presented with three counterfactual questions, one for each machine type. These were included to assess whether children accurately represented each individual machine's causal structure. Children were reminded that pressing a machine's button could activate one or more lights, which in turn could activate additional lights. They were shown an animation of a machine with two lights, illustrating how lights could control others. The animation also demonstrated that if one light was broken, it could not activate any other lights. To clarify this concept, the inside of the machine was displayed, showing the mechanism that allowed lights to turn each other on.

After watching the animation with two lights, children were again shown each of the three machine types used in the similarity judgment task, one at a time. For each, they were told one light was broken and asked to predict if a target light would turn on when the machine was activated. They first watched the machine operate without broken lights, then saw a red "X" over one light. Children made their choice by clicking on an image of either a lit or unlit lightbulb. They answered one counterfactual question for each causal structure.

Results

Similarity Judgments To analyze whether there were changes in performance based on age, we ran a generalized linear mixed effect model (GLMM) predicting children's similarity judgments (correct = 1, incorrect = 0) as a function of the following predictors: age (categorical variable including each age group from 4 to 7), trial type (common cause vs. common effect, common cause vs. causal chain, or common effect vs. causal chain), and subject as a nested random effect. GLMM was used due to the within-subject covariance structure, using the glmer function of the *lme4* package in R to specify fixed and random effects (Bates et al., 2015). The estimation method was maximum likelihood.

We find a main effect of age ($\chi2$ (3) = 20.89, $p < 0.001$), such that similarity judgments improve with age (4-year-olds: $M = .58$, 95% CI [.41, .75]; 5-year-olds: $M = .65$, 95% CI [.48, .81]; 6-year-olds: $M = .76$, 95% CI [.61, .90]; 7-year-olds: $M = .79$, 95% CI [.65, .93]). We also find that performance varied by comparison type ($\chi2$ (2) = 37.38, $p < 0.001$), and that these main effects were qualified by an interaction between age and comparison type ($\chi2$ (6) = 30.18, $p < 0.001$). This indicates that children in each age group did not perform equally well on all comparison types.

To follow up on the interaction between age group and comparison type, we examine performance against chance for each group. We compare to chance as we are interested in whether children at each age can make similarity judgments using causal structure, and not the relative performance of each age group. To account for multiple comparisons, we use a Bonferroni correction (corrected $\alpha = .05/12 = .0042$). Using one-sample *t*-tests, we found that 5-, 6- and 7-year-olds performed above chance for causal chain vs. common cause comparisons (all ps $< .004$), while 4-year-olds did not ($p = .801$). For causal chain vs. common effect comparisons, 6- and 7-year-olds (ps $< .004$), but not 4- and 5-year-olds ($p = .090$, $p = .007$, respectively), performed above chance. Finally, for common cause vs. common effect comparisons, none of the children performed differently than what would

be expected at chance (all *p*s > .09). This indicates that 6- and 7-year-olds only performed well on comparisons that included a causal chain machine, and that 4-year-olds failed to match based on causal structure for any of the machines.

Counterfactual Questions To check whether children were correctly interpreting the causal structure of the individual machines, we included counterfactual questions at the end of the experiment for each type of machine. Each child responded to one question for each causal structure, so we conducted a binomial test comparing whether children responded to this forced choice question differently that what would be expected at chance (.5) for each age group (correct = 1, incorrect = 0). We find that, for the common effect machine, no age groups responded different from chance (*p*s > .21). For the common cause machine, only 7-year-olds performed different from chance ($p = .007$; all other ages *p*s > .21). For the causal chain machine, only 6-year-olds performed different from chance ($p < .001$; all other *p*s > .11). Therefore, we find that children did not consistently respond to counterfactual questions accurately in this task.

Discussion

In Experiment 1, we find that 6-7-year-olds, but not 4-5-year-olds, match simple machines that share the same causal structure, which is consistent with past work using narratives (Rett et al., 2024). Critically, however, children of all ages were unable to recognize similar causal structures when comparing machines with a common cause to a common effect.

What might explain this difficulty? As detailed in the method, we carefully controlled the timing of light activations across machine types to ensure children could not succeed by simply tracking the time intervals between activations. For instance, for common cause machines, there was exactly one second between the activation of the first light and the activation of the final two lights. Similarly, for common effect machines, there was exactly one second between the activation of the first two lights and the activation of the final light. Since the timing between activations was identical for these two machine types, children could not determine similarity based solely on timing cues. In contrast, the causal chain machine had a different pattern: there was one second between the activation of the first and second lights, followed by another second before the third light activated. In this case, tracking the timing between events *could* help children correctly identify the causal chain machine. It is also possible that attentional lapses during the asynchronous presentation of the task may have interfered with recognition of causal structure, particularly for the counterfactual questions.

To address these concerns in Experiment 2, we move the study to a live, video chat format and modify the design to control temporal information across machines and disambiguate each machine's causal structure by showing children a single informative intervention on each.

Experiment 2

The preregistration for this experiment can be found at: https://aspredicted.org/kzhw-wq3k.pdf.

In Experiment 2, we introduce a novel set of machines with causal structures that 4-5-year-old children are able to infer from observing interventions: common cause (A ← B → C), common effect (conjunctive) (A→ C ← B), or causal chain (A → B → C) (Schulz et al., 2007; Lapidow & Walker, 2020). In contrast to Experiment 1, it is not possible to match any of the machines by relying on the timing of each event. Specifically, all events for all machines occur sequentially, with A happening first, B happening second, and C happening third. Children must therefore infer each machine's causal structure based on the observed intervention alone. For example, if preventing B does not prevent C, A must be the common cause of both B and C. Because children observed interventions on each individual machine, we did not include the counterfactual questions about hypothetical interventions that were used in Experiment 1.

Again, after showing children videos of each machine, we asked them to make similarity judgments about which machines in each set are "most alike." Given 4-5-year-olds' poor performance in Experiment 1, we modified the age range to include 6-9-year-olds.

Participants

A total of 128 6–9-year-olds participated, with 32 children in each age group (6-year-olds: M= 6.55 years, SD = .29; 7-year-olds: M = 7.47 years, SD = .27; 8-year-olds: M = 8.54 years, SD = .28; 9-year-olds: M = 9.47 years, SD = .31). This sample size was preregistered and is based on a power analysis for each age group's performance against chance. We calculated that with a sample size of 32 participants per group (128 participants in total) we would achieve at least 80% power to reject the null for effects as small as h = .5 (cohen's h, a measure of effect size used for comparing two proportions). An additional 19 children were tested but excluded due to parental or sibling interference (n = 9), technical difficulties (n = 4), experimenter error (n = 1), external distractions (n = 1), or failure to complete the study (n = 4). The study was conducted online, and participants were recruited via email through one of two pre-existing databases: a UC San Diego database of families in San Diego, CA and Children Helping Science's participant database.

Materials

The data and materials for Experiment 2 can be found at: https://osf.io/sqt7m/?view_only=298bc0dfbcd144dbade95232ab4d08b2.

A PowerPoint presentation was used to display the stimuli. A new set of animated, three-variable mechanical systems were created using Keynote (see Figure 2). There were six unique machines in total, with 3 functioning according to a common cause (A ← B → C), 2 functioning according to a common effect (conjunctive) (A→ C ← B), and 1

functioning according to a causal chain (A → B → C). We used this set of causal systems so that we could include 2 trials with unique causal structure comparisons: the common cause vs. causal chain, and the common cause vs. common effect. We include only these critical comparisons since they minimize other superficial cues (e.g. number of actions taken on each machine, and timing between each part's activation).

Each causal system was shown in a video, which included sound effects for each event. Unlike the light machines, the novel set of mechanical systems used in Experiment 2 do not contain low-level cues to similarity. That is, across machines, we control for perceptual features, such as color or shape, and all timing cues. The variables for each machine occur sequentially as a series of three events, occurring one at a time. To disambiguate causal structures, each video included an intervention on the machine. For example, to indicate that both effects produced by the common cause machine are independent of each other (B ← A → C), participants see an intervention on B, and observe no impact on of C.

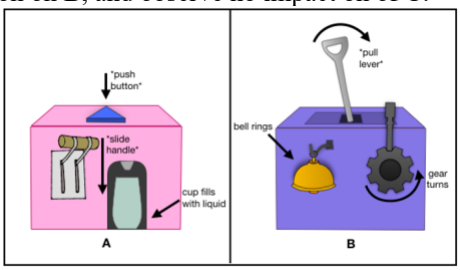

Figure 2. Example of two machines used in Exp. 2: a common effect (A) and a causal chain (B).

Procedure

Children were tested via Zoom. The experimenter shared their screen to display a PowerPoint presentation. Children were first introduced to a character, Bear (order of characters was counterbalanced). Children were told that Bear collected one kind of machine, and that the child would need to figure out which machine "went with" the one that Bear already had. The experimenter used the pronoun "they" when referring to each character, to prevent children from associating a particular color of machine with the character's gender.

Children then watched how the machine worked. For each machine, the display followed the same sequence. Children watched a video of how the machine worked twice, followed by 2 plays of a video showing a disambiguating intervention on the machine to establish the causal structure (i.e. stopping one part of the machine to demonstrate how the remaining parts work). After observing Bear's machine, children were asked, "What happened with the different parts of this machine?" This question ensured that children were engaged, without directing attention to any part in particular.

After viewing Bear's machine, children were shown two novel machines. They were told that they would have to find the machine that was the same kind as the one Bear already had. Children then saw videos of each of the novel machines, following the same sequence of events as they did for Bear's machine (a demonstration of the machine shown twice, an intervention shown twice, and a question asking what happened in the videos). After children were shown all three machines, they moved on to the test question. For the test question, children were asked "Which machine goes with the one that bear already has?" After selecting a machine, children were asked to explain their choice.

Children then proceeded to trial 2, which introduced a new character, Monkey. Trial 2 was identical to trial 1, but included a novel set of 3 machines that required children to compare different causal structures than those in trial 1. The order of trials was counterbalanced: In one trial, a common cause served as the "target" machine that children were prompted to match, and they had to select between a common cause and a causal chain. In the other trial, the "target" machine was a common effect, and children had to select between a common cause and common effect match. Trial order and machine location were counterbalanced across children, but machines within each trial remained the same (i.e. the same three machines were always viewed together).

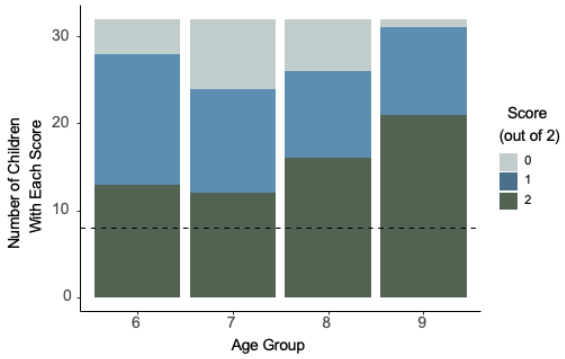

Figure 3. Number of children in Exp. 2 responding to the similarity judgment question at each score level (out of 2) by age group. Dotted line represents chance (.25).

Results & Discussion

Our main question was whether children matched machines based on their underlying causal structure. To analyze whether there were changes in performance based on age, we again ran a GLMM predicting children's similarity judgments (correct = 1, incorrect = 0) as a function of the following predictors: age (continuous), trial type (common effect or common cause target), the interaction between age and trial type, and subject as a nested random effect.

We found a significant main effect of age, Wald $\chi2(1) = 5.84$, $p = .016$, such that similarity judgments improved with age (6-year-olds: $M = .64$, 95% CI [.52, .76]; 7-year-olds: $M = .56$, 95% CI [.44, .68]; 8-year-olds: $M = .66$, 95% CI [.54, .77]; 9-year-olds: $M = .81$, 95% CI [.72, .91]). We also found a significant main effect of trial type, Wald $\chi2(1) = 10.03$, $p = .002$. There was no interaction between age and trial type, Wald $\chi2(1) = 3.01$, $p = .083$. As a result, we did not conduct separate comparisons at each age group. Instead, we conclude that, when collapsing age, children performed better when the target machine had a common effect structure ($M = .77$, 95% CI [.69, .84]) compared to when it had a common cause structure ($M = .57$, 95% CI [.48, .66]). That

is, the ability to match machines based on causal structure was dependent on the target machine's causal structure.

To analyze performance of each age group against chance, we compared the proportion of children who scored 2 out of 2 on the similarity judgment questions to a chance distribution of .25 using binomial tests. This method offers a more conservative measure of success than comparing the average score for each group to chance (.5). Instead, this allows us to set the criterion to the number of children who matched correctly on both items, as these children were the most likely to be relying on causal structure when making their similarity judgments. The results of these comparisons appear in Figure 3. The number of 8-, and 9-year-olds with a score of 2/2 was significantly higher than expected by chance (p = .003, p < .001, respectively). However, 6- and 7-year-olds' performance was not (p = .063, p = .106, respectively).[1]

These findings suggest that it is not until age 8-9 that children accurately match machines according to their shared causal structure in the absence of temporal cues. We also find differences in performance when comparing different machine types, which we address in the General Discussion.

General Discussion

Across two experiments, we find that children as young as 6 can categorize mechanical systems based on shared causal structure. However, until 8 years of age, children appear to base these inferences on low-level perceptual cues.

One possible explanation for this developmental trajectory is that younger children struggled to infer the causal structure of each individual machine based on intervention information alone. Although past work demonstrates that even 4-5-year-olds can discriminate causal structures after observing an informative action (Schulz et al., 2007), other work indicates that they struggle to do so when they are presented with conflicting temporal cues (McCormack et al., 2015). Given the importance of temporal information for inferring causal structure (McCormack et al., 2009; Lagnado & Sloman, 2004), younger children may be unable to discriminate causal structure in this task when the timing of variables is held constant.

Given this, what might explain children's success in inferring common effect categories over common cause categories? One possibility is that younger children relied on a different low-level cue—specifically, the number of actions taken on each machine (common effect machines required two actions, e.g., pushing a button and moving a slider, while common cause machines required one, e.g., pulling a lever). In a follow up study, we are examining whether children's understanding of the causal structure of each individual machine impacted their behavior in Experiment 2.

It is also possible that children correctly interpreted the machines' causal structure, but failed to recognize shared structure across machines. While even preschool-aged children represent their causal knowledge in terms of its underlying causal structure (Muentner & Bonawitz, 2017), the ability to successfully reason over these structures sometimes develops later. While some prior work shows that 4- to 6-year-olds can discriminate possible causal structures when presented with a forced choice between an informative and an uninformative action (e.g., Lapidow & Walker, 2020), other work suggests that children under 7 struggle to select interventions that disambiguate more complex three-variable systems (e.g., Frosch et al., 2012; McCormack et al., 2016; Meng et al., 2018). Thus, children's ability to use their knowledge of casual structure to match events may continue to develop over the course of middle childhood.

Future work will explore other cues that may facilitate children's recognition of shared causal structure. For example, it is possible that mechanistic information serves to bridge knowledge about a specific event (e.g., turning a car key powers the electrical system) with abstract understanding (e.g., a common cause structure linking the ignition, engine, and headlights; Keil & Lockhart, 2021). Exposing children to mechanistic details might therefore support their ability to recognize similar structural relationships across superficially distinct phenomena. Relatedly, it has been suggested that children struggle to give coherent answers about the effects of interventions when the underlying mechanism is unknown (Frosch et al., 2012). Although children can reason about causal relationships without detailed knowledge of mechanisms (e.g., recognizing that sunlight promotes plant growth without understanding the process of photosynthesis), it has been proposed that providing mechanistic information can facilitate the construction of more abstract knowledge (Chuey et al., 2021; Keil, 2022). Future work will investigate whether providing rich mechanistic details improve children's ability to categorize causal systems more broadly.

The current study also has several limitations that will be addressed in future work. First, real-world learning likely depends on the *usefulness* of abstracting common causal structure, and young children may be more likely to succeed when this information is expected to generalize to future contexts. Second, due to the relatively small number of causal machines used in Experiment 2, it difficult to assess whether children's pattern of responses reflect their failure to recognize common causal structure, or if it was due to stimuli-specific effects. Future work will include more (and more varied) exemplars to improve validity.

In conclusion, we find that, until age 8, children struggle to spontaneously recognize shared causal structure across mechanical systems. Given the necessity of making abstract connections across domains in early life, it is important to establish when children are capable of abstracting complex causal information across domains of knowledge.

Acknowledgements

Thank you to Iliana Kleiner, Sharon Lee, Mia Real, Crystal Coffey and Emily Songvilay for help with data

[1] A pilot study with undergraduates recruited via the SONA participant pool at UC San Diego (N = 40) indicated that adults performed near ceiling, with 85% of participants scoring 2/2 on similarity judgment questions (p < .001).

collection. We also thank the participating families. This work was supported by the Jacobs Foundation Fellowship and NSF CAREER (#2047581) to C. Walker and an NSERC PGS-D to A. Rett.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121-152.

Chuey, A., McCarthy, A., Lockhart, K., Trouche, E., Sheskin, M., & Keil, F. (2021). No guts, no glory: underestimating the benefits of providing children with mechanistic details. *npj Science of Learning*, *6*(1), 30.

Frosch, C. A., McCormack, T., Lagnado, D. A., & Burns, P. (2012). Are Causal Structure and Intervention Judgments Inextricably Linked? A Developmental Study. *Cognitive Science*, *36*(2), 261–285.

Galloway, K. R., Leung, M. W., & Flynn, A. B. (2018). A Comparison of How Undergraduates, Graduate Students, and Professors Organize Organic Chemistry Reactions. *Journal of Chemical Education*, *95*(3), 355–365.

Gelman, S. A., & Meyer, M. (2011). Child categorization. *WIREs Cognitive Science*, *2*(1), 95–105.

Goldman, S. R., Reyes, M., & Varnhagen, C. K. (1984). Understanding fables in first and second languages. *NABE Journal*, *8*(2), 35-66.

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, *137*, 137-153.

Keil, F. C. (2022). *Wonder: Childhood and the lifelong love of science*. MIT Press.

Keil, F. C., & Lockhart, K. L. (2021). Beyond cause: The development of clockwork cognition. *Current Directions in Psychological Science*, *30*(2), 167-173.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451.

Lapidow, E., & Walker, C. M. (2020). Informative experimentation in intuitive science: Children select and learn from their own causal interventions. *Cognition*, *201*, 104315–104315.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009). Cue competition effects and young children's causal and counterfactual inferences. *Developmental psychology*, *45*(6), 1563.

McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and Statistical Information in Causal Structure Learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(2), 395–416.

Meng, Y., Bramley, N., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In *Proceedings of the 40th annual conference of the Cognitive Science Society.*

Muentener, P., & Bonawitz, E. B. (2017). The development of causal reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 677–698). New York: Oxford University Press.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: *The National Academies Press*.

Nussenbaum, K., Cohen, A. O., Davis, Z. J., Halpern, D. J., Gureckis, T. M., & Hartley, C. A. (2020). Causal Information-Seeking Strategies Change Across Childhood and Adolescence. *Cognitive Science*, *44*(9).

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition*, 183, 57-66.

Nyhout, A., Henke, L., & Ganea, P. A. (2019). Children's counterfactual reasoning about causally overdetermined events. *Child development*, *90*(2), 610-622.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "Nearest Possible World." *Child Development.*, *81*(1), 376–389.

Rett, A., Amemiya, J., Hwang, B., Goldwater, M., & Walker, C. M. (2024). Children's recognition of causal system categories across superficially distinct events. *Developmental Psychology.*

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive science*, 36(5), 919-932.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental science*, 10(3), 322-332.

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

Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: Stages of expertise. *Journal of Research in Science Teaching : The Official Journal of the National Association for Research in Science Teaching*, *45*(7), 771–793.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266-281.

Williams, J. P., Lauer, K. D., Hall, K. M., Lord, K. M., Gugga, S. S., Bak, S. J., ... & deCani, J. S. (2002). Teaching elementary school students to identify story themes. *Journal of Educational Psychology*, *94*(2), 235.