Title

How to Change a Mind: Adults and Children Use the Causal Structure of Theory of Mind to Intervene on Others' Behaviors

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Wu, Shengyi
Schulz, Laura
Saxe, Rebecca

Publication Date

2024

Peer reviewed

How to Change a Mind: Adults and Children Use the Causal Structure of Theory of Mind to Intervene on Others' Behaviors

Shengyi Wu, Laura Schulz*, Rebecca Saxe*
{shengyiw, lschulz, saxe}@mit.edu
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139
* These authors contributed equally to this work.

Abstract

Prior studies of Theory of Mind have primarily asked observers to predict others' actions given their beliefs and desires, or to infer agents' beliefs and desires given observed actions. However, if Theory of Mind is genuinely a causal theory, people should also be able to plan interventions on others' mental states to change their behavior. The intuitive causal model of Theory of Mind predicts an asymmetry: one has to instill *both* the relevant belief and desire to cause an agent to act; however, to prevent a likely action, it suffices to remove *either* the relevant belief or desire. Here, we use these asymmetric causal interventions to probe the structure of Theory of Mind. In Experiments 1 and 2, both adults (N=80) and older children (N=42, 8-10 years) distinguished generative and preventative cases: selecting interventions on both mental states (both belief and desire) to induce an agent to act and just one of the mental states (either belief or desire) to prevent an action. However, younger children (N =42, 5-7 years) did not. To probe this age difference, in Experiment 3, we asked younger children (N=42, 5-7 years) just to predict the outcome of others' mental state interventions. Children predicted that interventions were more likely to prevent actions than to cause them, but failed to predict that intervening on both the relevant beliefs and desires is more likely to generate a novel action than intervening on either alone. These findings suggest that by eight to ten years old, people represent the causal structure of Theory of Mind and can selectively intervene on beliefs and desires to induce and prevent others' actions.

Keywords: Theory of Mind; causal interventions; social cognition; development

Introduction

As social creatures, humans' ability to navigate the complex social world is deeply intertwined with our capacity to understand and attribute mental states, such as beliefs and desires, to others. Scientists and philosophers have long considered that this capacity, known as Theory of Mind (ToM), is a causal theory that intentional agents' actions are caused by both their beliefs and desires (Dennett, 1987; Wellman & Bartsch, 1988; Gopnik & Meltzoff, 1997; Perner, 1991; Wellman & Woolley, 1990). Decades of research investigate how people predict others' behaviors based on their mental states, or infer their mental states based on observed behaviors (Baron-Cohen, Leslie, & Frith, 1985; Wellman & Bartsch, 1988; Flavell, 1999; Wimmer & Perner, 1983; Wellman, 2014; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Goodman et al., 2006; Gerstenberg & Tenenbaum, 2017). However, if Theory of Mind is a causal theory, people should also be able to use it to plan to change others' thoughts, and thereby their actions (Ho, Saxe, & Cushman, 2022).

In developmental research, the false-belief task is widely accepted as the standard measure of Theory of Mind. In false-belief tasks, children are asked to predict the action of an actor who holds wrong beliefs about the state of the world (e.g., where Sally will look for a marble that has been moved (Wimmer & Perner, 1983); what a person will think is inside a "Smarties" box which contains a pencil (Gopnik & Astington, 1988; Perner, Leekam, & Wimmer, 1987)). Children begin to pass these tasks between ages four and five, across task variations and cultures (Wellman, Cross, & Watson, 2001). Success in the false-belief task requires children to recognize that others' actions can be predicted and explained by their mental states.

However, the litmus test of whether children have a genuinely causal representation is their ability to use it for planning and intervention. We have known for several decades that even very young children can use their causal representations to intervene to both generate and prevent events, and screen off spuriously associated variables (Gopnik, Sobel, Schulz, & Glymour, 2001; Schulz & Gopnik, 2004). Children can also intervene selectively to change others' epistemic states (e.g. showing an agent something to change what they know (O'Neill, 1996)). To our knowledge, no prior work has tested whether children can plan interventions using their understanding of actions as jointly caused by relevant beliefs and desires.

Critically, the intuitive causal Theory of Mind model predicts an asymmetry of choice of interventions depending on whether the goal is to induce a new action or prevent a likely action. If the goal of interventions is to make someone perform an action that they would not otherwise take, the intervention must instill both the relevant belief and desire. We call this a generative intervention (See Figure1a). Taking an everyday example, if you want Anne to go to the post office, you have to make her both believe the post office is open and you have to make sure she wants to send something in the mail. On the other hand, if the goal is to prevent someone from taking an action that they would otherwise perform, a sufficient intervention would remove either the belief or desire. We call this a preventative intervention (See Figure1b). For instance, if Anne is planning to go to the post office and you want to prevent her from doing this, it suffices to change either just her belief (e.g., by telling her the post office is closed) *or* her desire (e.g., by mailing the package for her). These frameworks provide general causal structures for a ra-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

tional agent to plan to change others' actions by intervening on others' mental states.

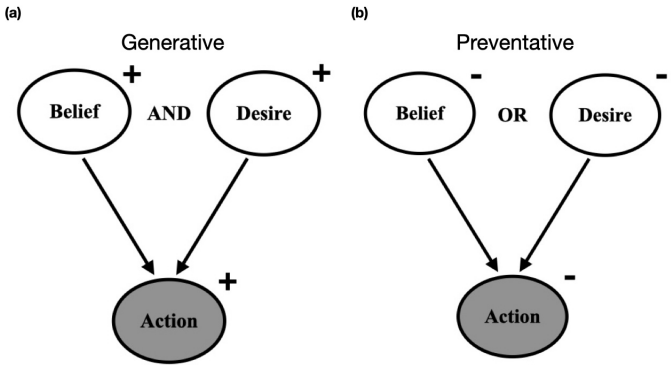

Figure 1: Causal Theory of Mind. If actions are caused by a belief and a desire, then (a) to induce a new action, one must instill both the relevant belief and desire, but (b) to prevent a likely action, it suffices to remove either the relevant belief or desire.

Here, we test whether people use an intuitive model of actions jointly caused by beliefs and desires to design interventions on others' actions, revealed by the signature asymmetry between generative and preventative interventions. In Experiment 1, adults watched animated story stimuli and chose interventions to achieve the goal of making an agent go to a new location (generative) or the goal of making an agent stop staying at the current location (preventative). In Experiment 2, children ages five to ten were tested with the same stimuli. To foreshadow, Experiments 1 and 2 reveal developmental change: adults and older children selectively choose appropriate interventions but younger children do not. Thus, in Experiment 3, we further probe the younger children's reasoning, by asking them to predict the result of a proposed intervention.

Experiment 1

Methods

Participants 104 adults were recruited via Prolific and paid $5 for completing the study. Adults were fluent English speakers from the United States, and gave informed consent. Twenty-four adult participants were excluded from analysis for failing any one of the three inclusion questions (n=23) or self-reporting at the end of the experiment that they did not understand the instructions (n=1). N = 80 adult participants were included in the analysis.

Materials and Procedures The stimuli consisted of animated videos created in Keynote and implemented online using jsPsych. Adults were told that the study was designed for children. The experiment began with a warm-up task, explaining that sometimes you only need to do one thing to make something happen (e.g., if you want someone to come to the door, you just need to ring the doorbell OR knock on the door. Doing both is not necessary) and sometimes you have to do two things to make something happen (e.g., if you want to bake cookies, you have to both turn the oven on AND put the cookies in. Doing only one of these is not enough).

The test stimuli consisted of two generative and two preventative stories presented in one of two counterbalanced orders (G, P, G, P or P, G, P, G). The assignment of stories to conditions was counterbalanced across participants. In each story, participants were introduced to a character (a monkey, mouse, rabbit, or bear) who had a preferred and a second-best snack option (e.g., the monkey likes grapes but LOVES bananas). In the generative condition, the goal was to make the character go to a target location (e.g., go to a chair); in the preventative condition, the goal was to make the character stop staying in the target location (e.g., get off the chair). There was also a non-target location (e.g., a basket) as well as other non-essential elements of the scene (e.g., a tire swing, trees, etc.) (See Figure 2).

The generative stories always began with the preferred snack in the non-target location. The intervention only on belief always took the form of putting the second-best snack on the target location so the character would know that it was there. The intervention only on desire always took the form of removing the preferred snack from the non-target location, so the character would want the second-best snack (See Figure 2A).

The preventative stories always began with the second-best snack in the target location. The intervention only on belief always took the form of removing the second-best snack from the target location so the character would know that it was absent. The intervention only on desire took the form of introducing the preferred snack to the non-target location so the character would not want the second-best snack anymore (See Figure 2B).

On the test trials, the participants were asked to choose between Bert, who always said the participant had to do two things (intervene on both the character's beliefs and desires) and Ernie, who always said that it was enough to just do one thing (either intervene only on the character's belief; or intervene only on the character's desire, counterbalanced across trials).

Results

If adults use an intuitive causal Theory of Mind to select interventions, they should intervene on *both* others' desires and beliefs more often in the generative than the preventative condition. To test this key prediction, we conducted a mixed-effects logistic regression predicting choices of intervention from condition, with random intercepts for subject and story. We coded Option "Bert" (interventions on both belief and desire are needed) as 1 if it was chosen and Option "Ernie" (intervention on either belief or desire is enough) as 0. We also contrast-coded the generative condition as 1 and the preventative condition as -1. There was a significant effect of condition ($\beta = 1.02, SE = 0.16, z = 6.240, p < .001$): adults chose to intervene on both belief and desire more of-

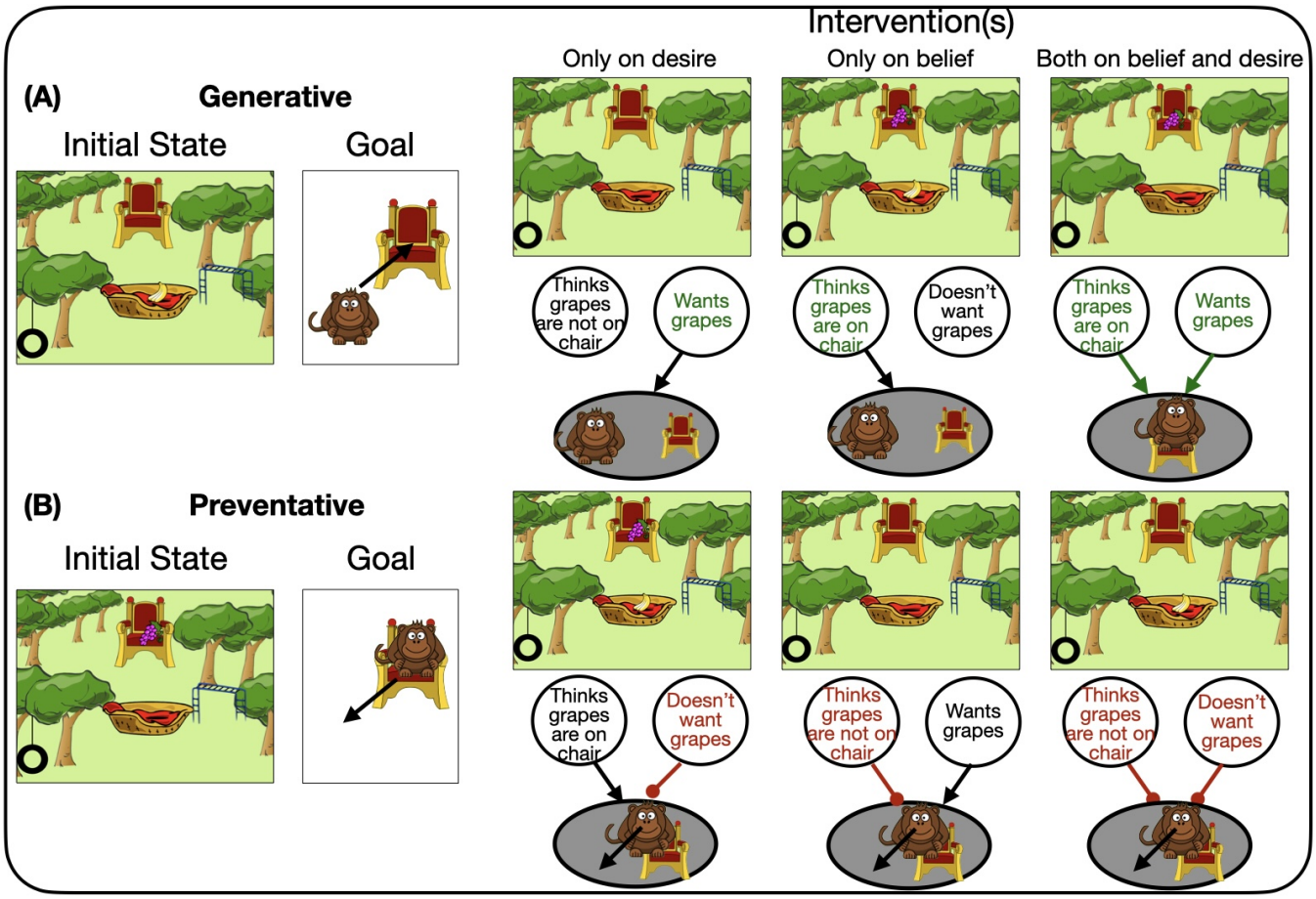

Figure 2: Design for Experiments 1 and 2. The colored texts show the change of mental state(s) after intervention(s).

ten in the generative condition (M = 0.89, SE = 0.03) than in the preventative condition (M = 0.53, SE = 0.08) (See Figure 3a). The intercept term was also significant ($\beta = 1.16, SE = 0.19, z = 5.87, p < .001$). Adults were more likely to choose to intervene on both mental states than only one mental state across all conditions.

We further explored adults' selection of interventions in the preventative condition. Adults were more likely to choose the option of intervening on both mental states when the intervention on one mental state was on belief (M = 0.64, SE = 0.03) than on desire (M = 0.42, SE = 0.03, $z = 2.631, p < .01$). That is, in the preventative condition, adults seem to expect that intervening only on desires is more likely to succeed than intervening only on beliefs.

Discussion

Consistent with the causal structure of an intuitive theory of mind, adults were more likely to intervene on both beliefs and desires to cause an agent to take a new action, but view an intervention on either belief or desire as more likely sufficient to prevent a planned action from happening. Note that adults selected the intervention on both beliefs and desires more frequently, across all conditions. Insofar as participants believe the interventions are low cost and only probabilistically effective, this is a rational decision: acting on both mental states maximizes the chance of achieving the desired outcome (Teo & Ong, 2023). Next, we tested whether children can systematically select interventions in the same task.

Experiment 2

Methods

Participants Child participants were recruited from and participated in the study asynchronously on the Children Helping Science website. Child participants were fluent English speakers, and tested with their caregivers' informed consent. Each eligible participant was given a $5 USD Amazon gift card.

The data were collected in two consecutive waves. First, fifty-nine 5-7-year-olds completed the study. Seventeen child participants were excluded from analysis for failing any one of the two inclusion questions (n=6), having technical difficulties (n=4), looking away from the screen or disappearing from videos for longer than 10 seconds consecutively (n=3), being ineligible for the study (e.g., ages, invalid accounts, n=4). N = 42 younger child participants (M = 6.52, range = 5.01-7.97 years) were included in the analysis.

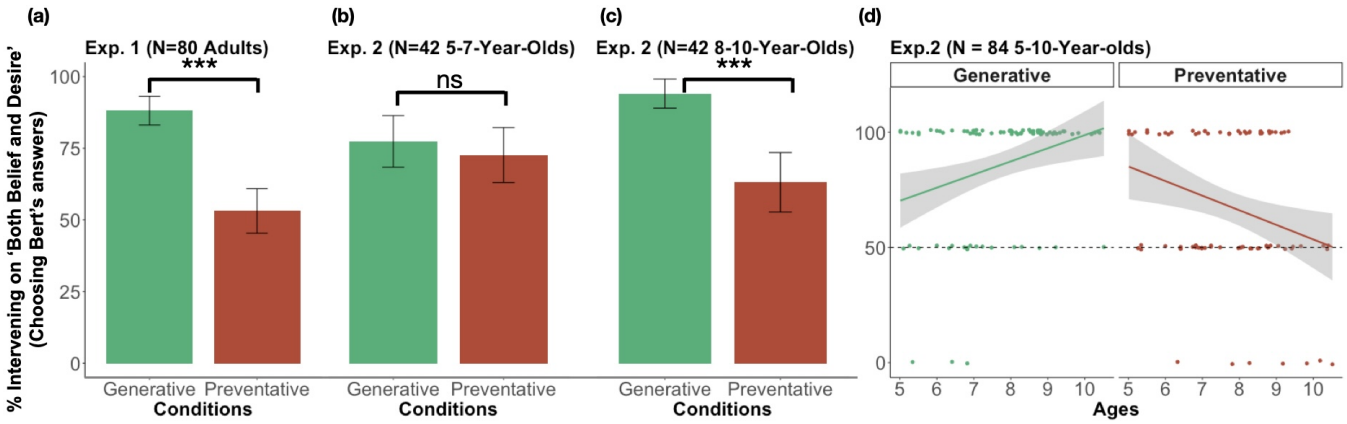

Figure 3: Percentage of choosing to intervene on both belief and desire in generative vs preventative conditions. (a-c) Error bars show 95% CIs around condition means. Both adults (a) and older children ages 8-10 (c), but not younger children ages 5-7 (b) selectively intervened on both mental states to induce an agent to act and just one of the mental states to prevent an action. (d) Probability of choosing to intervene on both belief and desire in generative vs preventative conditions, by age, for each child.

Second, fifty-one 8-10-year-olds completed the study. Nine child participants were excluded from analysis for failing any one of the two inclusion questions (n=5), having technical difficulties (n=1), looking away from the screen or disappearing from videos for longer than 10 seconds consecutively (n=1), parental interference (n=1), being not eligible for the study (n=1). N = 42 older child participants (M = 8.92, range = 8.00-10.52 years) were included in the analysis.

Materials and Procedures We used the same materials and procedure as in Experiment 1. The stimuli were implemented online on the Children Helping Science website. Child participants were included in the analysis if they answered the second and third practice trials correctly. Data analyses followed the same plan as Experiment 1.

Results

Results for 5-7-year-olds: There was no effect of condition on the selection of interventions ($\beta = 0.14, SE = 0.18, z = 0.74, p = .5$). Young children chose the intervention on both belief and desire equally often in the generative condition (M = 0.78, SE = 0.05) and in the preventative condition (M = 0.74, SE = 0.05) (See Figure 3b). Like adults, younger children chose to intervene on both mental states more often than only one mental state across conditions (intercept: $\beta = 1.19, SE = 0.23, z = 5.19, p < .001$);

We explored whether intervention selection changes with child age, by including a fixed effect of age (in months) in the model. This model did not explain significant additional variance (likelihood ratio test $\chi^2(1) = 0.084, p = .77$), and including an age by condition interaction did not improve model fit compared to the condition-only model (p = .24) or condition and age models (p = .10).

In the preventative condition, like adults, younger children were more likely to choose the option of intervening on both mental states when the intervention on one mental state was on belief (M = 0.83, SE = 0.003) than on desire (M = 0.62, SE = 0.01, $z = 2.09, p < .05$), suggesting that interventions on desire only were expected to be more effective.

Results for 8-10-year-olds: There was a significant effect of condition on the selection of interventions ($\beta = 1.15, SE = 0.27, z = 4.25, p < .001$). Older children were more likely to choose the intervention on both belief and desire in the generative condition (M = 0.94, SE = 0.01) than in the preventative condition (M = 0.63, SE = 0.06) (See Figure 3c). Older children were also more likely to choose the interventions on both mental states across conditions (intercept: $\beta = 1.73, SE = 0.35, z = 4.89, p < .001$). Adding age as a predictor significantly improved the model over the condition-only model (likelihood ratio test $\chi^2(1) = 5.43, p = .02$).

Like adults and younger children, in the preventative condition, older children were more likely to choose the interventions on both mental states when the intervention on one mental state was on belief (M = 0.82, SE = 0.02) than on desire (M = 0.45, SE = 0.04, $z = 2.97, p < .01$).

Combined results for 5-10-year-olds: To capture the potential developmental change, we combined data from 5-10-year-olds. The best-fitting model included an age by condition interaction; adding the interaction significantly improved the model fit compared to the condition-only model (likelihood ratio test $\chi^2(1) = 6.53, p = .01$) (See Figure 3d).

Discussion

In Experiment 2 older children (8-10 years old), but not younger children (5-7 years old), were more likely to intervene on both beliefs and desires to cause a novel action than to prevent an action. In the combined dataset, the condition by age interaction was significant.

Why did younger children not select different interventions to generate versus prevent actions? We know five-year-old children are able to predict others' actions based on their desires and (false) beliefs (Wellman et al., 2001). At least two

key differences between standard false-belief tasks and our novel intervention selection task may be relevant. First, in false-belief tasks children represent just one belief and one desire in the character, whereas in the intervention selection task, children must consider three sets of mental states: the character's current beliefs and desires, and the beliefs and desires the character will have after each intervention. Second, in false-belief tasks children are asked to make one prediction, whereas in the intervention selection task, children have to anticipate the consequences of two interventions in order to choose the best one. Either of these differences may increase the difficulty of intervention selection. To begin to disentangle these possibilities, in the next experiment, we test whether younger children (5-7 years old) *predict* different outcomes of interventions on just beliefs, just desires, or both.

Experiment 3

Methods

Participants Sixty-three 5-7-year-olds were recruited from and completed the study asynchronously on the Children Helping Science website. All child participants were fluent English speakers who did not participate in Experiment 2. Each eligible participant was given a $5 USD Amazon gift card. Twenty-one child participants were excluded from the analysis for failing both inclusion questions (n=2), having technical difficulties (n=2), looking away from the screen or disappearing from videos for longer than 10 seconds consecutively (n=8), or being ineligible for the study (e.g., ages, invalid accounts, n=9). N = 42 child participants (M = 6.38, range = 5.01-7.97 years) were included in the analysis.

Materials and Procedures Materials and procedures were similar to Experiment 2. Warm-up trials introduced the prediction task. In each of the two practice trials, participants predict ('yes' or 'no') on whether a certain event will happen when one potential cause occurs.

In each of the four test stories, participants were introduced to a central character (a monkey, mouse, rabbit, or bear) who had a preferred and a second-best snack option. At the end of each story, Ernie proposes to intervene on the animal character's mental state by changing the availability of snack(s), in order to generate or prevent an action (going to a target location). Across conditions, Ernie proposes to add or remove the preferred snack (to change the desire for the second-best snack), to add or remove the second-best snack (to change the belief about the location of the second-best snack), or both. Participants predicted whether Ernie's intervention would have its intended effect on the character's action. In the Generative condition, participants predicted whether the central character would go to the target location (e.g., go to a chair). In the Preventative condition, participants predicted whether the central character would stop staying at the target location (e.g., get off the chair). Participants gave a binary "yes" or "no" response to each question. The order of expected answers for the test trials (YNYN or NYNY) was counterbalanced across participants.

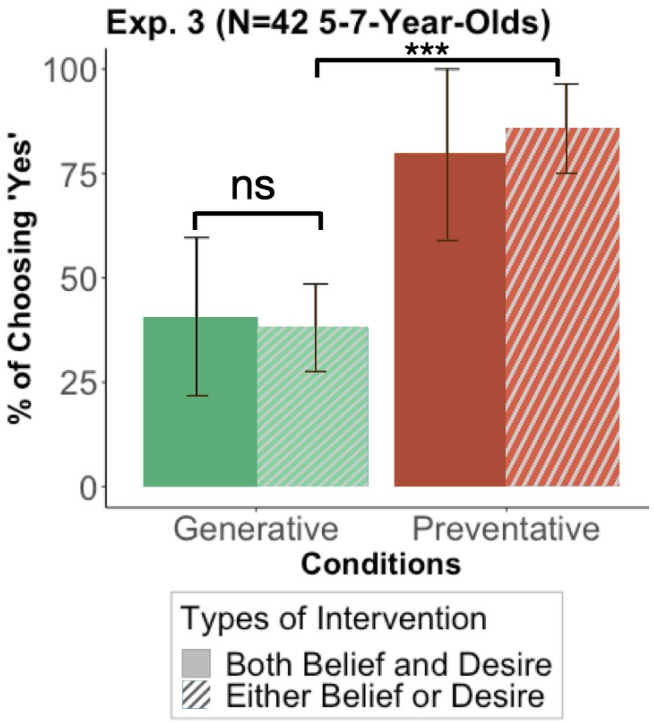

Figure 4: Percentage predicting "Yes", the intervention will achieve its aim, in generative vs preventative conditions, for interventions on both belief and desire, vs on either belief or desire

Results

Do younger children predict that a novel action will occur after an intervention on both belief and desire, but not after an intervention on either one? We conducted a mixed-effects logistic regression predicting choices of the predicted result of the intervention ('yes': 1 vs 'no': 0) from categorical types of intervention (generative&intervention on one mental state; generative&interventions on both mental states; preventative&intervention on one mental state; preventative&interventions on both mental states). The target action was not predicted to be more likely after interventions on both belief and desire, than after an intervention on either belief or desire ($\beta = -0.16, SE0.61, z = -0.26, p = .79$).

On the other hand, young children did predict that interventions to change either the belief or the desire would be more likely to prevent an action, than to generate an action ($\beta = 2.68, SE = 0.65, z = 4.12, p < .001$) (See Figure 4).

Discussion

In Experiment 3, we removed one challenging feature of the intervention selection task: instead of anticipating the consequences of two possible interventions in order to select one of them, children predicted the consequences of just one proposed intervention. Despite this simplification, 5-7-year-olds did not predict that interventions on both beliefs and desires were more likely to generate a novel action, than interven-

tions on either belief or desire alone. These results suggest that younger children's performance in Experiment 2 was not limited by the difficulty of selecting between interventions. Rather, the younger children may have a limited capacity to consider the consequences of simultaneously intervening on both a target character's belief and their desire. When the intervention would change only one mental state the children succeeded, correctly anticipating that interventions on either belief or desire alone are more likely to prevent an action, than to generate one.

General Discussion

We investigated the causal structure of people's intuitive theory of mind by asking people to select interventions to cause or prevent others' actions. Adults and older children systematically select interventions as predicted by a causal structure with two necessary causes: intervening on both beliefs and desires to generate actions, but intervening on either beliefs or desires to prevent actions.

The interventions considered here induce a voluntary change in the agent's behavior specifically by changing the belief and desire that are most relevant to causing the behavior. That is, we isolated interventions on the endogenous mental states to study the intuitive causal structure, at some cost to the naturalness of the scenario and interventions.

In the real world, of course, there are many other ways people might intervene to affect others' behavior. One could intervene physically (by moving someone to a location or blocking their path), or exogenously by introducing new costs or rewards (holding a gun to their head or offering a million dollars if they perform the action). And of course, perhaps the most obvious way to induce or prevent someone from acting is simply to ask: please go here or do not stay there. Moreover, people do not typically choose between intervening on either, or both, belief and desire. Most intuitive interventions on others' actions influence both desires and beliefs. If for instance, I tell you, "There is an ice cream truck at the park" my utterance is likely to – simultaneously – induce in you a desire for ice cream and a knowledge of where to get it. The single intervention will affect both mental states (and probably get you to the park).

None of these more quotidian interventions reveal whether people plan interventions using a model of actions jointly caused by beliefs and desires. Physical interventions do not test people's model of the mind at all. A verbal request to perform an action is a direct intervention on people's probability of acting, but requires no model of the typical mental causes of action.

One limitation of the experimental design is that interventions on beliefs and desires alone were not equated. At every age, participants expected interventions on desires to be more effective at prevention than interventions on beliefs. Because it is intuitively (and philosophically) hard to directly change others' desires except by changing associated beliefs, the intervention on desires used here introduced or removed a preferred competitor object. That is, the intervention manipulated whether the desire for the target object was causally relevant to the target action, without changing the characters' overall preferences (e.g., the monkey still preferred bananas to grapes, but wanted grapes if bananas were not available). When the bananas were in the basket, adults and children found it easy to predict that the monkey would not stay on the chair. By contrast, the preventative intervention on belief (remove the grapes) did not generate a specific contrasting action, and both adults and children were less confident it would be sufficient to prevent the monkey from staying on the chair. Future research should ideally manipulate the target belief and desire without generating a specific alternative action.

Note also that the intervention on desires did involve changing *some* of the agent's beliefs about the world (e.g., removing the bananas from the basket let the monkey know there were no bananas there and thus increased his desire for the grapes on the chair). For our purposes however, the critical point was that in giving the monkey information about the bananas, we did not change any of the monkey's beliefs about the grapes: the goal that ultimately motivated the target action. Future research could consider ways of intervening only on an agent's desires besides conveying information that changes the agent's (independent) beliefs about the world, for example, by satiating those desires.

Unlike older children and adults, younger children did not selectively choose interventions on both beliefs and desires, in order to generate actions. Indeed, younger children did not predict that a single proposed intervention on both beliefs and desires would successfully generate the target action. We speculate that 5-7-year-old children struggle to represent hypothetical simultaneous changes to both beliefs and desires.

Surprisingly, few prior studies have tested whether and when children can plan interventions using their Theory of Mind. The classic false-belief task, in which children must predict actions based on a character's desires and false beliefs, can be solved by non-causal predictive models, learned from observing sequences of actions (Rabinowitz et al., 2018). Unlike prediction tasks, planning requires a causal model that specifies the asymmetric dependence between causes and effects to identify the necessary and effective targets of intervention (Ho et al., 2022). The current results thus provide distinct, complementary support for the view of Theory of Mind as an abstract causal theory (Gopnik & Wellman, 1994).

Open Science

All experiments are preregistered. All preregistrations, stimuli, data, and code are available at this link: `https://osf.io/ka8dn/?view_only=3570d86146934062b27b31758ee1781e`

Acknowledgement

We thank Early Childhood Cognition Lab and Saxe Lab members at MIT for helpful discussion and feedback.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

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

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual review of psychology*, *50*(1), 21–45.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford University Press*.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the twenty-eighth annual conference of the cognitive science society* (Vol. 6).

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, 26–37.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, *37*(5), 620.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In *An earlier version of this chapter was presented at the society for research in child development meeting, 1991*.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child development*, *67*(2), 659–677.

Perner, J. (1991). *Understanding the representational mind.* The MIT Press.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, *5*(2), 125–137.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning* (pp. 4218–4227).

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental psychology*, *40*(2), 162.

Teo, D. W., & Ong, D. C. (2023). Instrumental causal learning. *PsyArXiv. October*, *27*.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition*, *30*(3), 239–277.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, *72*(3), 655–684.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275.

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