COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13163

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13163

What Could Go Wrong: Adults and Children Calibrate Predictions and Explanations of Others' Actions Based on Relative Reward and Danger

Nensi N. Gjata,[a] Tomer D. Ullman,[a] Elizabeth S. Spelke,[a] Shari Liu[b,c]

[a]*Department of Psychology, Harvard University*
[b]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*
[c]*Department of Psychological and Brain Sciences, Johns Hopkins University*

Received 14 June 2021; received in revised form 13 February 2022; accepted 9 May 2022

Abstract

When human adults make decisions (e.g., wearing a seat belt), we often consider the negative consequences that would ensue if our actions were to fail, even if we have never experienced such a failure. Do the same considerations guide our understanding of other people's decisions? In this paper, we investigated whether adults, who have many years of experience making such decisions, and 6- and 7-year-old children, who have less experience and are demonstrably worse at judging the consequences of their own actions, conceive others' actions as motivated both by reward (how good reaching one's intended goal would be), and by what we call "danger" (how badly one's action could end). In two preregistered experiments, we tested whether adults and 6- and 7-year-old children tailor their predictions and explanations of an agent's action choices to the specific degree of danger and reward entailed by each action. Across four different tasks, we found that children and adults expected others to negatively appraise dangerous situations and minimize the danger of their actions. Children's and adults' judgments varied systematically in accord with both the degree of danger the agent faced and the value the agent placed on the goal state it aimed to achieve. However, children did not calibrate their inferences about *how much* an agent valued the goal state of a successful action in accord with the degree of danger the action entailed, and adults calibrated these inferences more weakly than inferences concerning the agent's future action choices. These results suggest that from childhood, people use a degree of danger and reward to make quantitative, fine-grained explanations and predictions about other people's

Correspondence should be sent to Shari Liu, Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218-2686, USA. E-mail: shariliu@jhu.edu

behavior, consistent with computational models on theory of mind that contain continuous representations of other agents' action plans.

Keywords: Intuitive psychology; Cognitive development; Theory of mind; Danger

1. Introduction

Some actions are more dangerous than others: Skipping near the edge of a cliff is more dangerous than walking through a meadow, not because of the effort required to skip, but because of what could happen if you tripped. How do we understand dangerous actions when performed by others? In this paper, we explore whether adults and children are sensitive to the potential dangers other people face as they pursue their goals and whether adults and children expect others to quantitatively trade off potentially dangerous consequences of action failure against the potential rewards of success.

Using others' behavior to reason about their thoughts, beliefs, and goals, often termed intuitive psychology (Dennett, 1987), has long been a topic of study in cognitive science (Heider & Simmel, 1944). Recent computational proposals formalize this ability as a process of first assuming that others plan actions to maximize expected reward and minimize expected cost (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016):

$$U(A, S) = R(S) - C(A).$$

Given this forward plan, observers can then work backward from observed actions to the psychological causes of these actions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009). For example, the equation below allows the observer to infer the likely utility function (cost and reward) $P(R, C|A)$ that drives an agent's action by considering the prior likelihood of different cost and reward values, $P(C, R)$, and how well those values explain the observed action under a rational plan, $P(A|C, R)$.

$$P(R, C|A) \propto P(A|C, R) * P(C, R).$$

Past work suggests that even infants and young children can use these principles to reason about other people's minds and actions (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; Liu, Ullman, Tenenbaum, & Spelke, 2017). This has led researchers to propose that from early childhood, we model other people's minds and actions by (i) representing concepts like cost and reward as separable, abstract, continuous variables that are combined into plans which drive people's choices and (ii) inverting these plans to infer unobservable information about people's mental states.

Nevertheless, it is unclear how rich and systematic these abilities are in young children. Children are sensitive to absolute differences in utility when explaining and predicting other people's actions, but it is less clear whether they are also sensitive to relative differences in utility (e.g., Jara-Ettinger et al., 2015; Aboody, Zhou, & Jara-Ettinger, 2021; though see Bridgers, Jara-Ettinger, & Gweon, 2029): for example, are children more confident that someone will pursue a low cost or high reward action when the alternative option was much lower

in cost or much higher in reward, versus when the two options were almost equal in utility? Second, while children readily reason about the intended goals of other agents performing familiar actions (e.g., exploring toys, searching for objects), it is less clear whether children are able to apply this reasoning to agents and actions they themselves are unfamiliar with, and consider consequences of actions that neither they nor the actors observe. From one perspective, it is reasonable to propose that children cannot understand other agents' novel actions in such rich, systematic ways. Relative to adults, children's action planning abilities are less refined; they struggle to recognize the potential consequences of their actions (e.g., in judging the safety of situations like road crossings; Ampofo-Boateng & Thomson, 1991; Plumert et al., 2007), and for actions they are unfamiliar with (e.g., in judging the efficiency of grasping a novel tool; Ossmy et al., 2021). Nevertheless, this past research leaves open the possibility that children could express such abilities when reasoning about the actions of other people, at an age when they are beginning to take on more responsibility with their learning and decision-making (Rogoff et al., 1975), and when demands on their own motor planning are removed altogether. In summary, in order to fully test the model proposed by Baker, Jara-Ettinger, and colleagues and to ask whether such models include information about unobserved negative consequences, we sought to collect (i) quantitative data, (ii) from children and adults, and (iii) using stimuli that isolate for these negative consequences and that could plausibly be understood by children.

Thus, in the current research, we tested, first in adults (Experiment 1) and then in children (Experiment 2), whether continuous variables of reward (i.e., the positive utility of reaching the intended goal) and danger (i.e., the potential negative consequences if that plan fails) guide people's inferences and predictions in quantitative ways. We define danger as the negative utility associated with a possible state of the world if an agent's action were to fail (e.g., falling down a trench), reward as the magnitude of the positive utility associated with achieving a goal (e.g., crossing a dangerous-looking trench successfully to reach something on the other side), and physical cost as the effort required to carry out the action.[1.] In the current work, we focus on danger, operationalized as the height from which an agent could fall, for two reasons. First, even infants and non-human animals are sensitive to some aspects of danger in their own action planning. Studies of depth perception using "visual cliffs" in humans and other animals show that as infants learn to crawl and walk, they become sensitive to the depth of drop-offs on the ground plane and prefer to move towards shallower than deeper drop-offs (Gibson & Walk, 1960; Walk, Gibson, & Tighe, 1957). This sensitivity is quantitative: the greater the depth of the drop-off, the less infants are willing to climb down or reach their arms beyond its edge (Adolph, 2000; Kretch & Adolph, 2013). Studying danger in this way allows us to test whether children apply quantitative action plans to other people when they should have some specific knowledge about the affordances of the situation.

When people act in naturalistic contexts, their actions are driven by many correlated causes that are difficult to isolate and manipulate (e.g., dangerous actions are infrequent, correlated with proximity to the cause, and more likely to be chosen by some agents than others). Furthermore, naturalistic stimuli do not easily allow for tests of abstraction, which is central to the current research. Thus, research testing whether adults and children attribute continuous

abstract action plans to others often uses stimulus sets consisting of novel agents acting in simple environments (Baker et al., 2009, 2017; Gergely & Csibra, 2003; Jara-Ettinger et al., 2016). For the same reasons, we chose to use highly controlled stimuli inspired by classic visual cliff experiments, as well as more recent studies of infants (Liu et al., 2017) to test for quantitative judgments in adults (Experiment 1) and 6- and 7-year-old children (Experiment 2). Our stimuli consisted of a novel agent making decisions to jump over deep versus shallow trenches for goal objects. This stimulus set results in conditions that vary in danger and reward, controlling for the physical effort required to act and the probability of success. We tested whether under these conditions, adults and children can appreciate that (i) the agent acting could fall, that (ii) this fall could result in negative consequences, and (iii) the magnitude of these consequences varies with the height of the fall and therefore calibrate their predictions and inferences based on the degree of danger and reward they observe.

We found that both adults and children took into account consequences that they never observed when inferring an agent's preferences and predicting its future actions. Both children and adults tailored their judgments and predictions to the degree of danger and reward agents faced. However, both groups showed the smallest effect (in adults) or no effect (in children) when working backwards to infer the value of a goal from the specific degree of danger agents were willing to take in order to reach it. We also explored our results in two ways. First, we compared linear and non-linear statistical models of people's responses to explore the form of these judgments (see the Supporting Information for results). Second, we tested whether these trade-offs are best predicted by objective properties of the physical environment (i.e., the height that an agent could fall from), by people's expectations about how others appraise these situations (i.e., the agent's aversion to these heights), or both. From exploring the data, we found that expected appraisals added only a small amount or no explained variance to people's judgments, above and beyond the degree of danger imposed by objective metrics from the situation, suggesting that people were relying on the physical consequences of falling, rather than the agent's appraisals of these consequences, in making their judgments.

2. Experiment 1: Adults

Our experiment consisted of four tasks, designed to establish whether and how adults use danger to estimate, predict, and infer other people's actions and mental states. First, we asked to what degree people expect others to negatively appraise jumping over or falling into deeper trenches (estimation, Task 1a and 1b). Next, we examined whether people expect others to choose to jump over shallower trenches, all else being equal (prediction, Task 2). Finally, we asked whether people expect others to quantitatively trade off the potential negative consequences of jumping over deeper trenches against the potential reward of getting to goals on the other side (inference, Tasks 3 and 4). We first report information about our participants and the general procedure, and then cover the methods, analysis, and results from each task separately.

2.1. Methods

2.1.1. Participants

$N = 108$ adults (48 female, mean age = 39.86 years, range = 23–69 years) living in the United States were recruited through Amazon Mechanical Turk (mTurk) and were included in the final sample after exclusions. Fourteen participants were excluded for taking less than 4 min to complete the experiment ($n = 1$) or for failing a comprehension question or attention check ($n = 13$). Our sample size was based on a power analysis from Task 3, which demonstrated the weakest effect in a pilot study (see Figs. B1 and B2 for pilot data in the Supporting Information). The minimum suggested sample size for adults was 40 for adults and 30 for children. Because of the relative ease of collecting data from adult participants, we chose a substantially larger sample size. We did not collect demographic information beyond gender and age, and past work suggests that the racial demographics of the mTurk participant pool are similar to those in the general U.S. population (Burnham, Le, & Piedmont, 2018). All data collection methods and procedures were approved by the Committee on the Use of Human Subjects at Harvard University. Participants took a median of 10.66 min to complete the study. The sample size, participant inclusion criteria, methods, and data analysis plan for this study were pre-registered on the Open Science Framework. All materials are available at https://osf.io/u8b9s/.

2.1.2. Materials and design

This four-part experiment was deployed on Qualtrics. Our stimuli were adapted from events shown to infants in previous work from our lab (Liu et al., 2017). The stimuli featured an agent (a red sphere with eyes and a smiling mouth), reward objects of varying color and shape (e.g., cones, cylinders, and prisms, etc), and seven trenches of constant width (2 units in Blender space) but varying depth (1–7 units in Blender space). We fixed trench width to control for the physical cost associated with jumping across the trenches. All participants saw one set of objects for each task, with no overlap in color-object combinations between tasks. There were two versions of each object set per task that consisted of different order pairings between trials and objects.

The order of tasks 1, 2, and 3 was counterbalanced using a Latin Square (for visualizations of data from individual tasks broken down by task order, see Figs. B14 and B15 in the Supporting Information). Task 4 always appeared last due to the concern that the language within each trial ("the deepest cliff [the agent] would jump") could influence participants' judgments in the other tasks, although see the Supporting Information for results from an additional experiment showing that seeing Task 4 prior to versus after Task 3 (its closest analog) does not appear to change the results of either task. Participants never saw the agent fall, but they were asked to consider the agent's mental state in the hypothetical situation if it were to fall in Task 1 (which was randomly assigned to appear first, second, or third across participants).

2.1.3. Procedure

After giving consent to participate, people were introduced to an agent who could jump over different trenches to reach objects on the other side and saw a video of a ball rolling into a shallow, medium, and deep trench to convey that this animated world has normal physics. Next, they completed questions on judging the relative depth of trenches, and on their uncertainty about the agent's initial dispositions towards the trenches ("Before she acts, do we know which trench [the agent] wants to jump?"—the correct answer of "no") and the objects ("Before she acts, do we know which things she likes?"—the correct answer of "no"). Before each of the four tasks, participants answered three questions about the relevant continuous measure (e.g., for judgments about the agent's preference, "Where would you put the slider if you think [the agent] likes the object a little bit?"). The comprehension questions that preceded each task were designed during an extensive piloting and validation phase of the research to make sure that children understood the task and how to provide a continuous response. Because we wanted to keep the methodology similar between adults and children, both children and adults underwent the same procedure. If participants answered a comprehension question incorrectly, they were prompted to re-read the question and try again (adults) or were given feedback about why their answer was wrong and prompted to try responding again (children). To make the task more child-friendly, we referred to the agent as "Wendi" or "she," rather than "it" or "the agent," but throughout the paper, we use the terms "the agent" and "it," following the conventions of previous literature.

2.1.4. Data and analysis

We used linear mixed-effects models (Bates, Mächler, Bolker, & Walker, 2015) in R for all analyses. While we originally pre-registered our analyses to account for the maximum random effects structure of the data (Barr, Levy, Scheepers, & Tily, 2013), allowing each participant a random slope and a random intercept, we pared down some models in the main analysis to only include a random intercept for participants due to model convergence issues.[2] Our significance threshold was a two-tailed alpha level of 0.05. All reported p-values are two-tailed, and all degrees of freedom were generated using Satterthwaite's method.

We also conducted two exploratory analyses, asking (1) whether linear or non-linear models better accounted for the data in Tasks 1–4 (see the Supporting Information) and (2) how people's individual assessments of danger are related to their predictions and inferences in the other tasks, (2a) how well people's judgments of how the agent feels about each situation in Task 1a and 1b predict their responses when the agent faced the very same trenches in Task 2 (predicting action) and Task 3 (inferring reward),[3] and (2b) cross-task correlations between how strongly people's responses tracked with our manipulations. We used Akaike Information Criteria to assess model fit and parsimony. In the main text, we report the main findings from these exploratory analyses. (For full details, see the Supporting Information.)

2.2. Task 1: Do people associate deeper trenches with more negative utility?

First, we asked whether participants expect others to ascribe negative utility to the trenches in the current stimulus set. This served as a way of validating our primary task manipulation.

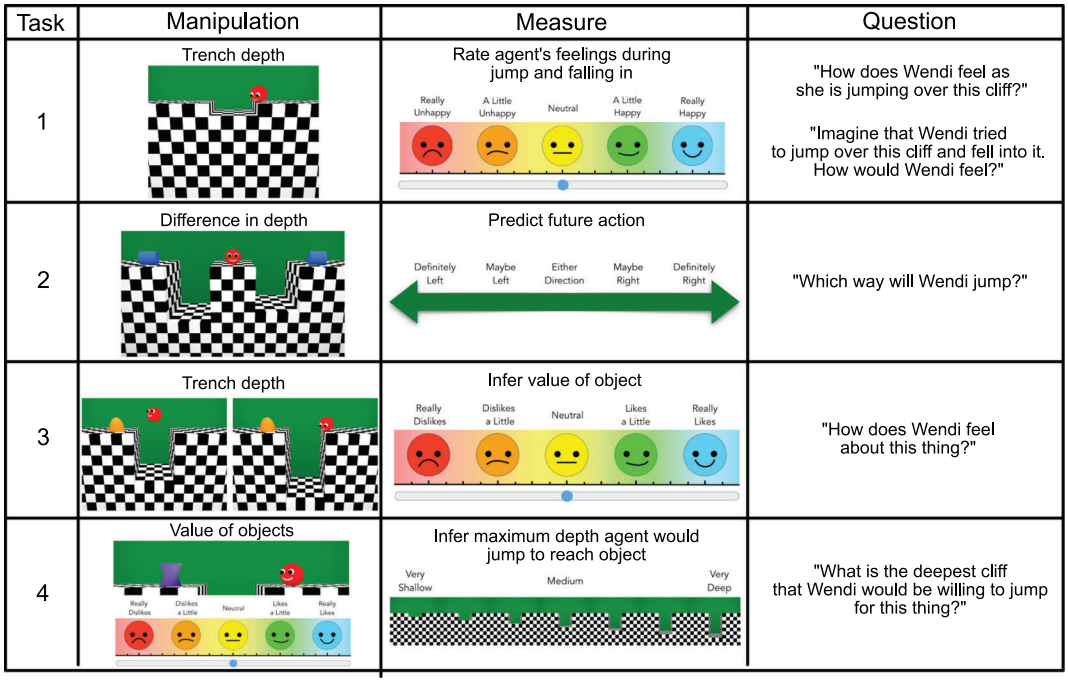

Fig. 1. Overview of the four tasks in Experiments 1 and 2, including the main manipulations and measures, and the verbatim question presented to participants.

2.2.1. *Methods*

Participants saw images of the agent facing trenches of varying depth (1 unit to 7 units in Blender space) in random order. Participants then rated how the agent would feel as it was jumping (Task 1a), and how it would feel if it fell in (Task 1b), on a scale that ranged from "really unhappy" (0) to "neutral" (50) to "really happy" (100). The left-right anchors of the scales were counterbalanced between participants and consistent within participants. (See Fig. 1.)

2.2.2. *Results*

We found that as the trenches became deeper, people judged that the agent felt more negatively, both as it was jumping, 95% confidence interval (CI) $[-7.299, -4.765]$, unstandardized $B = -6.032$, standardized $\beta = -0.474$, standard error $(SE) = 0.644$, $t(107.012) = -9.374$, $p < .001$, and if it were to fall in, $[-5.784, -4.910]$, $B = -5.347$, $\beta = -0.561$, $SE = 0.223$, $t(646.143) = -23.99$, $p < .001$ (pre-registered, confirmatory analysis). These two ratings were correlated with each other, $[0.445, 0.552]$, $r(751) = .500$, $p < .001$ (exploratory analysis), and people rated the agent's emotions as more negative in situations where it fell ($M = 23.713$, $SD = 19.091$) than while the agent was jumping ($M = 52.086$, $SD = 25.455$), $[-29.991, -26.707]$, $t(752) = -33.898$, $p < .001$ (exploratory analysis). This finding shows

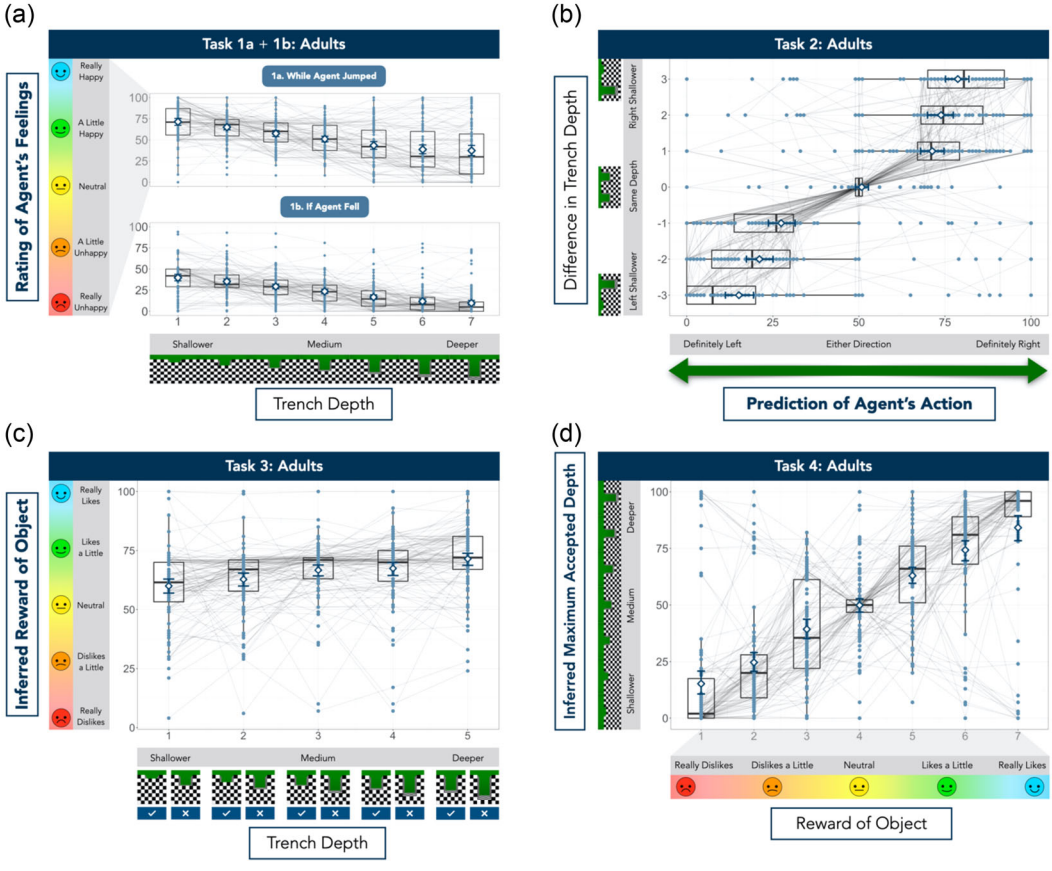

Fig. 2. Results from Tasks 1–4 in Experiment 1 ($N = 108$ adult participants). Bold axis names indicate dependent measures (the vertical axis for Tasks 1, 3, and 4 and the horizontal axis for Task 2). Connected individual points indicate data from a single participant, diamonds and error bars indicate means and bootstrapped 95% CI around the mean, and boxes indicate the middle two quartiles of data. (a) People rated that the agent would feel worse while jumping over deeper trenches (1a) and if it fell into deeper trenches (1b). (b) People predicted that the agent would jump the shallower trench and became more certain as the difference in depth between the two trenches increased. (c) People's ratings for how highly the agent values the reward object tracked with the depth of the trench the agent willingly jumped for that object. (d) People's inferences for how deep a trench the agent would jump for an object tracked with how much the agent reported liking that object.

that participants expected the agent to feel worse about deeper trenches, both before acting, and if its actions were to fail. (See Fig. 2a.)

2.3. Task 2: Do people use danger to predict the actions of others?

In Task 2, we tested whether adults appreciate that danger can influence another person's future actions and expect others to minimize danger, holding equal the physical cost of actions and the rewards these actions lead to.

2.3.1. Methods

In each trial, participants saw an agent facing a choice to jump one of two trenches to reach one of two identical goal objects. Our main manipulation was the *difference in depth* between the two trenches: One trench remained fixed at a medium depth (4 Blender units) while the other randomly varied between 7 depths, ranging from shallow (1 unit) to deep (7 units). In relative terms, this creates a depth difference from −3 (variable trench was 3 units shallower than the fixed trench) to +3 (variable trench was 3 units deeper). Whether the left or right trench varied in depth was counterbalanced across participants and consistent among participants. Each scenario featured a different pair of identical goal objects.

Across seven trials, participants used a sliding scale to indicate which direction they thought the agent would jump, ranging from "definitely left" to "definitely right" with "either direction" as the midpoint.

2.3.2. Results

We found that as the relative depth between trenches increased, people were more likely to judge that the agent would jump the shallower trench, [10.691, 13.580], $B = 12.148$, $\beta = 0.776$, $SE = 0.732$, $t(107.011) = 16.61$, $p < .001$ (pre-registered, confirmatory analysis). These results indicate that people used the magnitude of the difference in depth between the two trenches in order to make a prediction about which trench the agent would jump over. (See Fig. 2b.)

In exploratory analyses, we found that after controlling for the objective depth of the trenches, there was a small effect of people's ratings for how the agent would feel with respect to these trenches (from Task 1a and 1b) on their responses in Task 2. For example, the worse people thought the agent would feel while jumping over the deep trench, relative to the shallow trench in Task 1, the more they thought that the agent would choose the shallow trench in Task 2 ([−0.189, −0.020], $B = -0.105$, $\beta = -0.073$, $SE = 0.043$, $t(749) = -2.409$, $p = .016$). We found a similar marginal effect for people's ratings for how the agent would feel if it fell, [−0.197, 0.011], $B = -0.093$, $\beta = -0.055$, $SE = 0.053$, $t(749) = -1.755$, $p = .080$.

2.4. Task 3: Do people reason that the choice of a more dangerous action implies that the goal of the action brings a higher reward?

The two previous tasks are basic tests of the broad hypothesis that people indeed see these situations as dangerous and expect others to avoid danger. In Tasks 3 and 4, we tested for a deeper inference: Do adults infer the rewards of goals from the danger that others were willing to withstand for those goals, and do these inferences about value vary quantitatively with manipulations of danger?

2.4.1. Methods

In Task 3, we varied how deep a trench the agent was willing to jump for a goal object. First, participants saw animations of an agent jumping over a trench, and backing away from the trench, as examples of what it means for the agent to accept versus reject each jump. In each trial, participants saw two images: one showing the agent jumping a trench to reach

an object and one showing the agent declining to jump a trench 2 units deeper for the same object. Across five trials, participants rated how much the agent valued that object on a continuous sliding scale that ranged from "really like" (0) to "really dislike" (100) with "neutral" (50) as the midpoint. Which anchor appeared on the left versus right was consistent within participants across trials and counterbalanced between participants. Trials were shown in random order.

2.4.2. Results

We found that people's ratings for how much the agent valued the goal object varied with how deep a trench the agent previously jumped for that goal object. With each increasing unit of depth, the rating of the agent's liking increased by 2.761, or 0.266 standard deviations, on average (full-range 0–100), [1.993, 3.530], $B = 2.761$, $\beta = 0.266$, $SE = 0.390$, t(107)=7.072, $p < .001$ (pre-registered, confirmatory analysis). This result shows that people use the amount of danger others face when pursuing their goals to infer the value of these goals. (See Fig. 2c.)

In exploratory analyses, we found that people's ratings in Task 1a ([−0.083, 0.046], $B = -0.018$, $\beta = -0.027$, $SE = 0.033$, $t(532.890) = -0.553$, $p = .580$) and 1b ([−0.031, 0.130], $B = 0.050$, $\beta = 0.061$, $SE = 0.041$, $t(529.706) = 1.204$, $p = .229$) did not predict their value judgments in Task 3, above and beyond the objective depths of the trenches we presented. (See the Supporting Information for details.)

2.5. Task 4: Do people reason that others are willing to face greater dangers for more valuable goals?

In Task 4, we asked whether people perform the inference from Task 3 in the opposite direction: Do people expect others to withstand more danger for more highly valued goals?

2.5.1. Methods

We manipulated how much the agent reported liking a new set of objects, using the same scale from Task 3. Across seven trials, the agent reported valuing the object at seven uniformly spaced levels (from "really dislikes" (1) to "really likes" (7)). Participants were then asked to rate on a sliding scale the deepest trench the agent would be willing to jump for the object, given how much the agent likes it (full scale: 0–100). (See Fig. 1.) The left-right anchors of the preference scale were consistent within participants across Tasks 3 and 4 but counterbalanced across participants. Trials were presented in random order.

2.5.2. Results

We found that people's judgments about trench depth varied depending on how highly the agent valued the goal objects on the other side. With each added point on the liking scale (1–7), people judged that the agent would jump a trench 11 units deeper to reach them (full range 0–100), [10.945, 12.609], $B = 11.777$, $\beta = 0.710$, $SE = 0.424$, $t(646.766) = 27.772$, $p < .001$ (pre-registered, confirmatory analysis). This finding suggests that people use the

value of goals to estimate the amount of danger others would face to obtain these goals. (See Fig. 2d.)

2.6. Discussion

In Experiment 1, we found that adult participants expected others to negatively appraise (Task 1a) and feel worse if their actions failed (Task 1b) when in more dangerous situations, expected others to minimize danger (Task 2), and expected others to trade off the danger of actions and the rewards they bring (Tasks 3 and 4). Adults' ratings are tracked with objective features of the physical situation (how far the agent would fall if its actions were to fail), more so than their estimates of how bad that outcome would be for the agent.

3. Experiment 2: Children

Next, we asked whether these representations of reward and danger are present in the social reasoning of 6- and 7-year-old children. Here we ask whether such representations support graded judgments, which are difficult to measure in younger children but nevertheless are a key signature of a rich and productive system for reasoning about other minds (e.g., not just inferring what others prefer, but how much; not just predicting what others will choose, but at what level of certainty). Compared to adults, children have less experience making high-stake decisions, and under some conditions do not distinguish between naturalistic safe versus dangerous situations like traffic crossings (Ampofo-Boateng & Thomson, 1991; Plumert et al., 2007). Do they, nevertheless, hold rich abstract and continuous knowledge of people's actions and plans, when demands on their own motor planning are removed? In a second pre-registered experiment, we ask whether children, like adults, make quantitative trade-offs between danger and reward in order to predict and explain other people's behaviors.

3.1. Methods

3.1.1. Participants

$N = 36$ children (24 female, mean age = 7.01 years, range = 6.02–7.83 years) were included in the reported analyses and recruited from the greater Boston area to participate at the Harvard Lab for Developmental Studies. The demographics of this sample roughly matched with those typically from our lab (majority White, with parents who have at least a college degree). No children met the pre-registered exclusionary criteria, so all 36 children were included in the final sample. Because we were interested in whether judgments of danger guide rich, quantitative predictions, and inferences about other people's actions and minds, we chose to focus studying children at an age when they are beginning to take on substantial responsibility for their own learning and decision-making (Rogoff et al., 1975), and at an age where children, in previous studies and during our pilot experiments, could understand and use continuous scale measures to report their responses (Gweon & Asaba, 2018). As in Experiment 1, we chose this sample size based on a simulation power analysis over pilot data from Task 3 (see Fig. B2 in the Supporting Information for pilot findings).

3.1.2. Procedure

This research was conducted on the cusp of the COVID-19 pandemic, and so we tested 24 children in our lab and 12 children online, using video conferencing (see Chuey et al., 2021, for an overview of methods for online testing). Children saw the same survey and measures as adults, presented with more child-friendly instructions by an experimenter. We adapted methods for testing young children online (SocialLearningLab, 2020) to ask for consent, conduct an audio-visual setup, and debrief the participants and families. Legal guardians gave consent for their children to participate, and all children gave assent. Families received travel compensation (in-lab participants only) and a small thank-you prize (all participants) for participating. Study sessions typically lasted about 45 min in the lab and 55 min online.

For materials, datasets, data analysis files, and pre-registration see https://osf.io/u8b9s/.

3.2. Results

3.2.1. Task 1 results

Children predicted that the agent would feel worse when faced with a deeper trench, both as it was jumping (Task 1a), $[-8.679, -5.400]$, $B = -7.040$, $\beta = -0.530$, $SE = 0.826$, $t(35.003) = -8.524$, $p < .001$ and if it fell in (Task 1b), $[-8.200, -5.429]$, $B = -6.815$, $\beta = -0.556$, $SE = 0.711$, $t(40.618) = -9.589$, $p < .001$ (pre-registered, confirmatory analysis). (See Fig. 3a.) Like adults, children's ratings across these two measures were correlated, $[0.403, 0.589]$, $r(250) = 0.502$, $p < .001$, and children rated the agent would feel worse if it fell ($M = 24.480$, $SD = 24.554$), compared to while it was jumping ($M = 53.306$, $SD = 26.638$), $[-32.003, -25.647]$, $t(251) = -17.863$, $p < .001$ (both exploratory analyses).

3.2.2. Task 2 results

Children predicted that the agent would choose to jump the shallower trench, and their predictions were stronger as the deeper trench grew in depth, $[10.560, 14.904]$, $B = 12.732$, $\beta = 0.740$, $SE = 1.094$, $t(35) = 11.64$, $p < .001$ (pre-registered, confirmatory analysis). (See Fig. 3b.) Exploratory analyses showed that in contrast to adults, controlling for the main task manipulation, children's ratings from Task 1, for how the agent felt about each trench as it was jumping (Task 1a) ($[-0.096, 0.205]$, $B = 0.055$, $\beta = 0.040$, $SE = 0.077$, $t(247) = 0.709$, $p = .479$) and if it fell in (Task 1b) ($[-0.046, 0.256]$, $B = 0.105$, $\beta = 0.075$, $SE = 0.077$, $t(247) = 1.353$, $p = 0.177$) did not reliably predict their judgments in Task 2.

3.2.3. Task 3 results

In contrast with adults, children's ratings for how much the agent valued the goal object did not vary with the depth of the trench the agent jumped for that object, $[-3.271, 0.260]$, $B = -1.506$, $\beta = -0.088$, $SE = 0.896$, $t(105.662) = -1.68$, $p = .096$ (pre-registered, confirmatory analysis). (See Fig. 3C.) Children's ratings differed significantly from those of adults, 95% CI $[-5.753, -2.781]$, $B = -4.267$, $\beta = -0.334$, $SE = 0.758$, $t(574) = -5.628$, $p < .001$ (exploratory analysis)

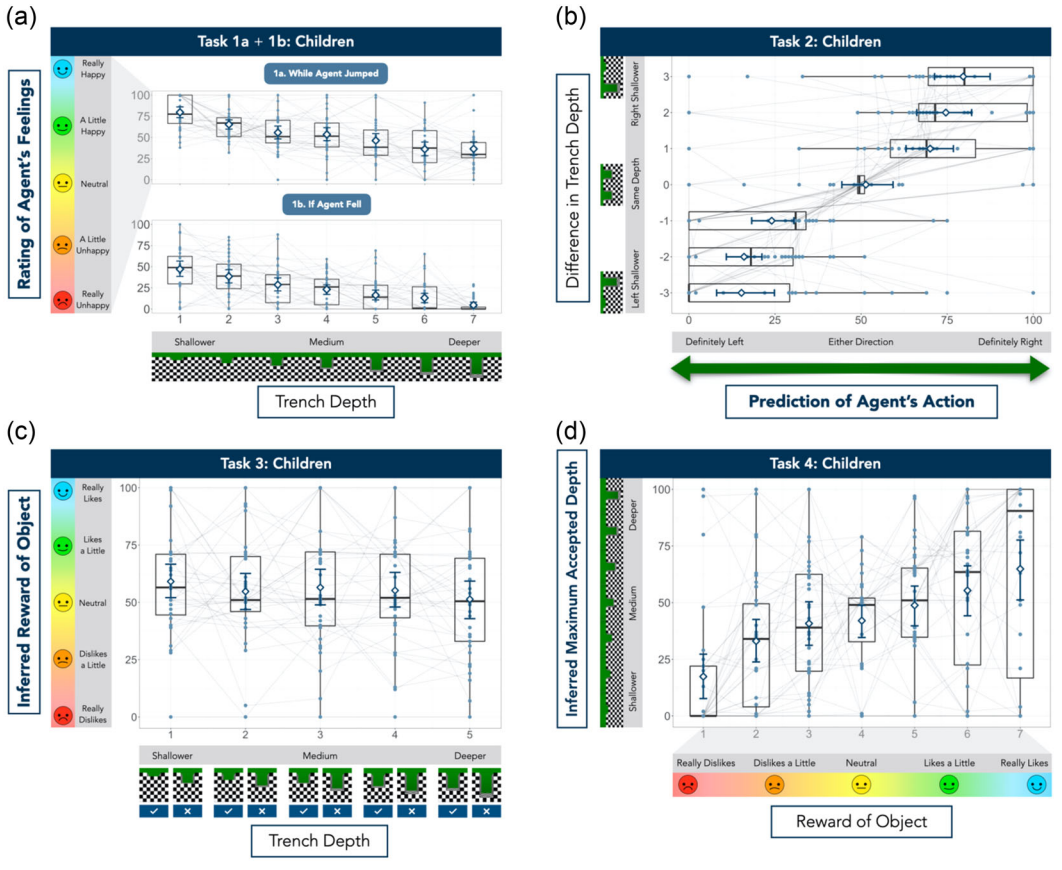

Fig. 3. Results from Experiment 2 ($N = 36$, 6- and 7-year-old children). Bold axis names indicate dependent measures. Connected individual points indicate data from a single participant, diamonds and error bars indicate means and bootstrapped 95% CI around the mean, and boxes indicate the middle two quartiles of data. (a) Children rated that the agent would feel worse while jumping over (1a) and if it fell into (1b) deeper trenches. (b) Children predicted that the agent would jump the shallower trench and became more certain as the difference in depth between the two trenches increased. (c) Children's judgments of how highly the agent values the reward object did not track with the depth of the trench the agent willingly jumped for that object. (d) Children's predictions for how deep a trench the agent would jump for an object tracked with how much the agent reported liking that object.

3.2.4. Task 4 results

Like adults, children predicted that the deepest cliff the agent would be willing to jump for an object scaled with how much the agent reported liking the object, [3.356, 10.577], $B = 6.966$, $\beta = 0.408$, $SE = 1.819$, $t(34.999) = 3.83$, $p < .001$. (See Fig. 3d.)

3.2.5. Across-task relationships for both children and adults

The exploratory analyses above examined how individual people's responses to *specific trenches* in Task 1 relate to predictions and explanations about the same trenches in Tasks

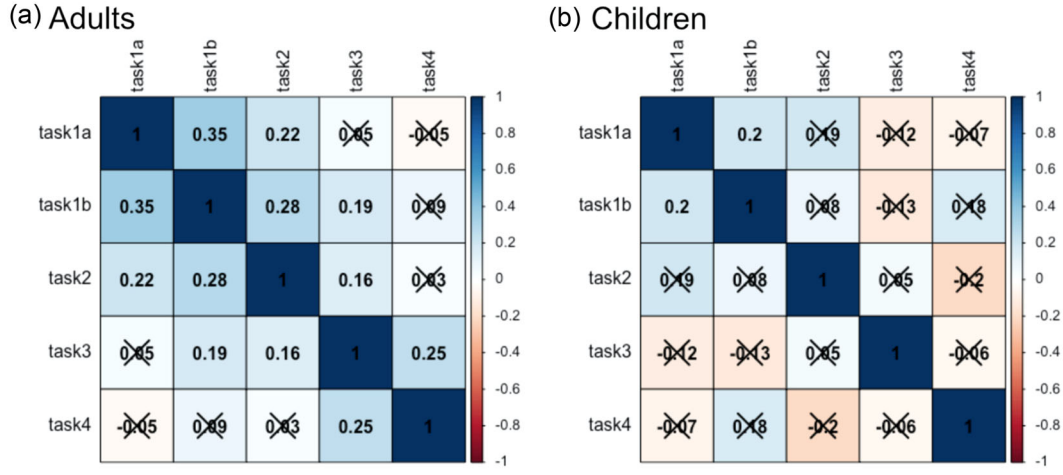

Fig. 4. Correlation matrices across Tasks 1–4 for (a) adults and (b) children. The numerical value in each cell indicates Kendall's tau (τ) between participants' z-scored responses across two tasks. The hue and saturation of each cell indicate the direction and strength of this relationship. X-marks indicate non-significant relationships across tasks (alpha = 0.05).

2 and 3. To examine whether people's responses to our manipulations of reward and danger were related more broadly across tasks, we conducted the following analysis. First, for each participant, we computed Kendall's tau (τ) for how their responses scaled with our main manipulation (trench depth in Tasks 1–3 and reward in Task 4). For each task, we then z-scored these τ values, such that participants with higher scores showed a bigger effect for that task, relative to other participants. Then, for each participant, we generated a meta-correlation coefficient (also using Kendall's τ), which represents people's pairwise performance across tasks. Figure 4 reports these results for both children and adults. We found that, while adults' responses for each task related to their responses for at least one other task, the only significant cross-task relation that we found in children was their ratings of how the agent would feel as it was jumping versus if it fell into trenches of different depths.

3.3. Discussion

In Experiment 2, we found that 6- and 7-year-old children, like adults, expect others to negatively appraise potentially dangerous situations (Task 1a) and feel worse if their actions failed in these situations (Task 1b), expect others to minimize danger (Task 2), and expect others to withstand greater danger for objects they value more highly (Task 4). In Task 3, children and adults diverged: Whereas adults inferred that the more danger someone was willing to withstand for a goal, the more the person values that goal, children did not respond systematically to this task. We also found that in comparison to adults, children's responses across tasks were less coherent. This could be because children's responses are noisier or less reliable in general, or because we were underpowered to detect such effects using our current sample. Overall, we found, in three out of four tasks, that both adults and children expect

other agents to take into account the continuous rewards of goal states and how the amount of danger is involved in reaching those goal states, holding constant physical effort.

4. General discussion

Recent computational proposals of theory of mind hypothesize that when people reason about the minds and actions of other people, we appeal to continuous variables like the strengths of people's preferences and beliefs, as well as the costs imposed by their physical constraints, and use this information to infer the causes of action and predict future behavior (Baker et al., 2009, 2017; Jara-Ettinger, Schulz, & Tenenbaum, 2020). In the current research, we tested whether the plans that adults and children attribute to other people (1) take into account consequences that neither they nor the actor ever observes and (2) whether these consequences guide quantitative judgments about what other people want and will do. Using a set of stimuli that are less familiar to adults and children than everyday situations, but that children could plausibly understand, we systematically varied the reward of an agent's goal states (i.e., how much the agent likes or dislikes different objects) and the dangerous consequences of its actions (i.e., if the agent were to fall, how far they would fall).

Our confirmatory analyses showed that adults and children expect others to negatively appraise dangerous situations (Task 1), predict that others will choose safer over more dangerous actions (Task 2), and expect others to trade off how badly actions could end against the reward of successfully reaching the intended goal (Task 4). We found that adults (but not children), under this task setting, systematically reasoned that the greater the amount of danger someone faces to reach a goal, the more they value that goal (Task 3). Our exploratory analyses suggested that both adults and children relied more on the height an agent could fall from than on the agent's appraisal of this fall above and beyond this height, in order to make their judgments. Their appraisal ratings explained very little or no additional variance in their explanations and predictions, and there were only moderate cross-task correlations in the data from adults. Nevertheless, we note that our study was not set up to evaluate individual differences, or to tease apart the contributions of two highly correlated predictors (height of fall and appraisal of this fall), so the results of these exploratory analyses should be interpreted with caution. Altogether, our findings show that people are tuned into the degree of danger and reward that others face and trade off these variables to explain and predict the actions of others. We also highlight that the materials and methods presented in this paper, all available at https://osf.io/u8b9s/, could be adapted to study-related questions in the theory of mind in childhood, some of which we cover below.

4.1. An intuitive theory of psychology that includes concepts of danger

Our results broadly support and extend the framework theory that people understand the minds and actions of others by assuming that others plan their actions (Bayesian theory of mind; Baker et al., 2009, 2017), taking into account variables like cost and reward (the naive utility calculus; Jara-Ettinger et al., 2016). First, our results show that the utility we ascribe to

others' decisions goes beyond weighing the negative cost of acting and the positive rewards that those actions lead to (Jara-Ettinger et al., 2016). Our findings suggest that in addition, people are sensitive to aspects of action that cannot be picked out by any particular path features, but rather depend on the potential negative consequences of acting. (Indeed, all the actions that participants viewed involved identical action trajectories, and the stimuli from each task featured just still images from these trajectories.) We argue that in order to succeed at this task, participants had to understand that (i) the agent acting could fall, that (ii) this fall results in negative consequences, and (iii) the magnitude of these consequences varies with the height of the fall.

Our data are consistent with at least two possible conceptions of danger, and, currently, do not adjudicate between them. First, people could represent danger, $D(A)$, as a negative reward, that trades off, like physical cost $C(A)$, against the positive reward of goal states, $R(S)$, and results in a utility of that action-state pair, $U(A, S)$:

$$U(A, S) = R(S) - C(A) - D(A),$$

whereas cost describes the physical work associated with action, danger picks out an additional negative value of that action independent of physical work. Here, the danger is defined over actions (e.g., jumping over a deep trench), with no explicit representation of the resulting state itself.

Alternatively, people may represent the multiple possible states, $S_i \in S$, that an action may generate, including the positive rewards associated with achieving goal states, and the negative rewards associated with failing to do so (e.g., in our case study, falling). The expected value of an action, under this account, depends on the probability of transitioning to each of these states, $P(S_i|A)$, the reward associated with each state, $R(S_i)$, and the cost of the action needed to make this transition, $C(A)$:

$$U(A, S) = \sum_{S_i \in S} P(S_i|A)R(S_i) - C(A).$$

This second conception of danger explicitly relies on hypothetical representations of possible futures, or counterfactual past states that did not occur (but could have), and the probabilities that these futures could or could have become real. For now, it remains an open question which of these two models is a better description of people's conception of danger.

It is also unclear what specific psychological and physical knowledge supports judgments of danger and the origins of these abilities in development. Do adults and children define danger over the physical environment (e.g., falling from a greater height is worse), an agent's bodily sensations (e.g., longer falls are worse because they lead to more severe injury or more pain), or an agent's mental states (e.g., longer falls are worse because they elicit more fear and require more courage to overcome)? Furthermore, it is unknown whether adults and children only consider the negative states that result from bad outcomes or also the projected positive cost of this fallout. For example, in the situations we studied here, do adults and children consider what happens after the agent falls? Falls from greater heights could result in more injury, but they also require greater physical effort to get out of deeper trenches, and these efforts are less likely to succeed, potentially leading to further injury. While the results from Task 1

show that children and adults expect others to negatively appraise dangerous situations, we found that individual differences in these judgments explained little to none of the individual differences in people's explanations and predictions, after controlling for physical features of the environments people saw. We note, though, that our task was not designed to sensitively measure such differences. Future work, focusing on specific queries about the agent's psychological states, physical properties, and the content of its hypothetical or counterfactual states (e.g., how risk-seeking or risk-averse people think this agent is, how afraid, thrilled, or bored the agent feels, what would happen to the agent's physical body if it fell, or what the agent would do after it fell) could reveal how our conceptions of people as mental and physical beings enter into our understanding of the dangers we and others face. Lastly, while infants begin to avoid steep drop-offs in the second year of life (Kretch & Adolph, 2013), it is less clear whether infants make use of negative consequences to understand the actions of other agents. Ongoing work is currently testing these abilities, using the same stimulus set, in infants (Liu, Ullman, Tenenbaum, & Spelke, 2019).

4.2. Inferring value from danger: How do adults and children compare?

There are several possible explanations for why adults systematically inferred the value of a goal from the danger a person faces to reach it, but children did not.

As a reminder, in Task 3, participants saw an agent jump over a trench to reach a goal object, refuse to jump a slightly deeper trench to reach that same goal, and then rated how much the agent valued that object. It is possible children and adults interpreted the main question for Task 3 ("How does she feel about this thing?") differently. Children could have interpreted our question to mean, "How does [this agent] feel about this situation more broadly?," "How does [this agent] feel when jumping over this cliff?," or "Given how difficult or easy it was to reach this [neutral-looking] toy, does the agent feel this toy was worth the danger?" Any of these interpretations of our question could have resulted in responses that pair deeper cliffs with lower values, which we did observe in some children (see Fig. 3). However, it is not clear to us why such explanations apply to this particular task and not others. It is also possible that children's attention to frequency information (the agent always accepts one trench and refuses another) outweighed their attention to the changing depths of these trenches across trials.

Although adults reasoned that more dangerous actions to reach goals indicate higher goal value, this effect was the smallest we observed ($ß = 0.266$ standard deviations, or an increase in 2.7 out of 100 points of inferred value for every added Blender unit of trench depth). Thus, we need to account for why this effect was absent in children and was small in adults. Briefly, it is possible that for both children and adults (i) inverting someone's plan is more difficult or noisy than merely running it forward (or, that counterfactual reasoning is more difficult or noisy than hypothetical reasoning; Beck, Robinson, Carroll, & Apperly, 2006) and/or that (ii) reasoning about value is fundamentally an overdetermined problem, with many possible causes, for both adults and children.[4.] Here, we asked whether people can infer how highly other agents value different goal objects based on what they are willing to do (i.e., how much danger they are willing to endure), but participants could have considered other factors, like the features of the object, what participants themselves prefer, and why particular

objects are paired with particular obstacles. (This may explain why children, on average but not reliably, paired deeper cliffs with lower-value objects: Perhapsthe object was placed there intentionally, to discourage the agent from moving to it.) Children's hypothesis space for why people value the things they do may differ from that of adults; for example, children may be willing to entertain a wider array of event features as explanations for what others like and how much. In sum, it is possible that for children in particular and for people more generally, other people's stated reward functions constrain our predictions about what they will do, more so than their actions constrain our inferences about how much reward they place over their goals, because the latter requires inverting their action plans, and/or because their reward functions are overdetermined.

4.3. Limitations

There are several key limitations to this work. First, both adult and child participants were convenience samples and the children we studied in particular were not representative of the U.S. population with respect to race and socioeconomic status. Thus, for now, our inferences over these results should be constrained to the population we studied. Furthermore, it is unclear whether these fine-grained representations of reward and danger guide people's responses during tasks that do not explicitly ask them to make these judgments. Lastly, by virtue of using a highly controlled stimulus set, we traded ecological validity for experimental control, to enable us to address our question of interest. As a result, this research does not address whether adults or children would similarly calibrate their responses in more naturalistic contexts.

4.4. Conclusion

In conclusion, this work shows that starting in childhood, we understand other people's behaviors by considering not only the goals of their actions but also the consequences they did not intend. Furthermore, children's and adults' judgments showed that they not only take into account these consequences but take them into account in rich quantitative ways, consistent with the broader proposal that our theory of other people's minds and behaviors depend on abstract, continuous representations of their action plans.

Acknowledgments

We gratefully acknowledge the following funding sources: the Center for Brains, Minds, and Machines (CBMM), funded by NSF Science and Technology Center award CCF-1231216, the Harvard College Research Program (to NG), NSF (Graduate Research Fellowship under grant DGE-1144152 to SL), and NIH (Ruth L. Kirschstein National Research Service Award under F32HD103363 to SL). Thank you also to Rebecca Saxe and the members of the Harvard Lab for Developmental Studies for feedback and support, and to our participants and their families.

Conflicts of interest

The authors have no conflicts of interest to declare.

Notes

1. We do not address a related concept, *risk* (the probability of achieving the intended outcome; Kahneman & Tversky, 1979), in this work, though we consider how danger may rely on probability in the discussion.
2. This happened for the confirmatory analysis in Tasks 1b and 4 for adults only. All exploratory models were pared down to include a random intercept for participants for simplicity.
3. Because we did not manipulate trench depth in Task 4, this trial-specific analysis was only possible to conduct in Tasks 2 and 3.
4. We do not believe the fixed order of the two tasks (Task 3, then Task 4) drove this effect the Supporting Information, Fig. B16 for data from an additional sample of 34 adult participants who performed these two tasks in counterbalanced order.

Author contributions

NG and SL conceptualized research and developed methodology with feedback from TU and ES. NG validated methods, designed stimuli and other study materials, and collected data with support from SL TU, and ES. ES provided lab resources necessary to carry out work. SL supervised the project. NG organized, visualized, analyzed, and archived data and other materials. NG and SL wrote the original draft of the paper, and TU and ES contributed to revising and editing.

Open Research Badges

The experiments reported in this paper were formally pre-registered on the Open Science Framework at https://osf.io/gm87c. In the results, we explicitly mark which analyses are confirmatory versus exploratory following this plan, and all derivations from the plan. All materials for the work presented in this paper (data, analysis scripts, and protocols) are available at https://osf.io/u8b9s/.

References

Aboody, R., Zhou, C., & Jara-Ettinger, J. (2021). In pursuit of knowledge: Preschoolers expect agents to weigh information gain and information cost when deciding whether to explore. *Child Development*, *92*(5), 1919–1931.

Adolph, K. E. (2000). Specificity of learning: Why infants fall over a veritable cliff. *Psychological Science*, *11*(4), 290–295.

Ampofo-Boateng, K., & Thomson, J. A. (1991). Children's perception of safety and danger on the road. *British Journal of Psychology*, *82*(Pt 4), 487–505.

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*(March), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

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

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.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, *4*, 144–152.

Burnham, M. J., Le, Y. K., & Piedmont, R. L. (2018). Who is mTurk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion, and Culture*, *21*(9–10), 934–944.

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers of Psychology*, *12*, 4968.

Dennett, D. C. (1987). *The intentional stance*. London: MIT Press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Science*, *7*(7), 287–292.

Gibson, E. J., & Walk, R. D. (1960). The "visual cliff." *Scientific American*, *202*(4), 64.

Gweon, H., & Asaba, M. (2018). Order matters: Children's evaluation of underinformative teachers depends on context. *Child Development*, *89*(3), e278–e292.

Heider, F., & Simmel, M. (1944). An experimental study of social behavior. *American Journal of Psychology*, *57*(2), 243–259.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Science*, *20*(8), 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naïve utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kretch, K. S., & Adolph, K. E. (2013). Cliff or step? posture-specific learning at the edge of a drop-off. *Child Development*, *84*(1), 226–240.

Liu, S., Ullman, T., Tenenbaum, J., & Spelke, E. (2019). *Dangerous ground: Thirteen-month-old infants are sensitive to peril in other people's actions*. PsyArXiv. https://scite.ai/reports/10.31234/osf.io/rvydk

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Ossmy, O., Han, D., Kaplan, B. E., Xu, M., Bianco, C., Mukamel, R., & Adolph, K. E. (2021). Children do not distinguish efficient from inefficient actions during observation. *Scientific Reports*, *11*(1), 18106.

Plumert, J. M., Kearney, J. K., & Cremer, J. F. (2007). Children's road crossing: A window into Perceptual-Motor development. *Current Directions in Psychological Science*, *16*(5), 255–258.

Rogoff, B., Sellers, M. J., Pirrotta, S., Fox, N., & White, S. H. (1975). Age of assignment of roles and responsibilities to children. *Human Development*, *18*(5), 353–369.

Social Learning Lab (2020). Online testing: Startup guide and materials, https://doi.org/10.5281/zenodo.3762737

Walk, R. D., Gibson, E. J., & Tighe, T. J. (1957). Behavior of light- and dark-reared rats on a visual cliff. *Science*, *126*(3263), 80–81.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting information