

# Look before you leap: Quantitative tradeoffs between peril and reward in action understanding

Nensi N. Gjata<sup>1</sup> (nensi\_gjata@college.harvard.edu)

Tomer D. Ullman<sup>1,2</sup> (tullman@fas.harvard.edu)

Elizabeth S. Spelke<sup>1,2</sup> (spelke@wjh.harvard.edu)

Shari Liu<sup>1,2</sup> (shariliu01@g.harvard.edu)

<sup>1</sup>Department of Psychology, Harvard University

<sup>2</sup>Center for Brains Minds and Machines, MIT

Cambridge, MA 02143 USA

## Abstract

When we reason about the goals of others, how do we balance the positive outcomes that actions led to, with the potentially bad ways those actions *could* have ended? In a four-part experiment, we tested whether and how adults (full study) and 6- to 8-year-old children (ongoing study) expect other agents to take account of the ways their goal-directed action could have failed. Across 4 different tasks, we found that adults expected others to negatively appraise perilous situations (deep trenches), to minimize the danger of their actions, and to trade off danger and reward in their action plans. Our preliminary children's study shows similar trends. These results suggest that people appeal to peril—how badly things could go if one's actions fail—when explaining and predicting other people's actions, and also make quantitative inferences that are finely tuned to the degree of peril and reward that others face.

**Keywords:** intuitive psychology; cognitive development; theory of mind

## Introduction

Some actions are more dangerous than others: Walking near a the edge of a deep trench is more perilous than walking through a sunny meadow, not because of the energy required to walk, but because of what could happen if the person were to trip. How do we as observers understand these actions when performed by others? In this paper, we explore whether adults and children are sensitive to the potential dangers other people face while performing actions, and whether they expect others to quantitatively trade off the negative consequences that a dangerous action could produce against the rewards of the goal states that the action aims to achieve.

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 assuming that others plan actions to maximize expected utility by maximizing reward and minimizing cost (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), and working backwards from their actions to work out hidden causes, like beliefs and goals (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009). The ability to use actions to reason about minds is available even to young children and infants (Liu, Ullman, Tenenbaum, & Spelke, 2017; Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015), suggesting that it is an early-emerging and central part of our social intelligence.

In this paper, we focus on "peril" or "danger" as a potential variable in our intuitive psychology. We define peril as the magnitude of a negative alternative 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 value associated with achieving a goal (e.g. crossing a dangerous-looking trench successfully to reach something on the other side), and cost as the physical effort required to carry out the action <sup>1</sup>.

We focus on danger and peril for two reasons. First, there is strong reason to think that even infants and non-human animals are sensitive to some aspects of peril in their action planning. Studies of depth perception using "visual cliffs" on humans and other animals show that as soon as infants gain the ability to walk or crawl, they become sensitive to the depth differences of surfaces, and prefer to move to shallower than deeper surfaces (Gibson & Walk, 1960; Walk, Gibson, & Tighe, 1957). Moreover, this sensitivity is quantitative: the larger the distance between surfaces, the less likely infants are willing to climb down or reach their arms beyond the edge of a cliff (Kretch & Adolph, 2013; Adolph, 2000).

Second, human infants appear to be sensitive to peril even when analyzing other people's actions. Inspired by studies of infants' decisions to navigate visual cliffs and real cliffs, recent research indicates that infants are capable of using danger to predict and explain others' behaviors. In ongoing work (Liu, Ullman, Tenenbaum, & Spelke, under revision), we found that Thirteen-month-old infants looked longer in surprise when an agent jumped a deep trench towards a goal when they could have jumped a shallower trench instead, and inferred an agent's preferences based on how deep a trench the agent previously jumped for its goals. These findings suggest that peril, at least in the context of heights and cliffs, is important not only for motor planning but also for understanding other people's actions.

Here, we rely on the methods from the classic visual cliff experiments and more recent infant looking time experiments to test for more fine-grained, quantitative judgments in adults (full study) and 6- to 8-year-old children (ongoing study). To test whether people quantitatively integrate danger and reward to reason about others' plans, we used continuous mea-

<sup>1</sup>We do not address a related concept, *risk* (the probability of the intended outcome; Kahneman & Tversky, 1979), in this work, though we speculate how representations of danger may rely on information about probability in the discussion

asures to ask whether people expect others to systematically trade off danger and reward when inferring an agent’s preferences and predicting its future actions. We also explored our results in two ways. First, we compared linear and non-linear models of people’s responses to explore the form of people’s judgements. Second, we tested whether these trade-offs are best predicted by objective properties of the physical environment (i.e. the actual depth of the trenches people saw), by people’s expectations about how others appraise these situations, or both.

### Experiment 1: Adults

Our experiment consisted of 4 tasks. In our baseline task, we ask to what degree people expect others to be negatively impacted by jumping over or falling into deeper trenches (estimation, Task 1). Next, we examine whether people expect others to choose jumping over shallower cliffs, all else being equal (prediction, Task 2). Finally, we ask whether people expected others to quantitatively trade off the negative potential peril of jumping over deeper trenches against the potential reward of getting to a goal 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.




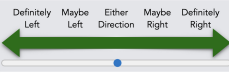



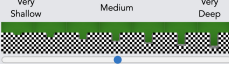
Task	Manipulation	Measure
1		Rate Agent’s Feelings While She Jumps and If She Fell In 
2		Predict Future Action 
3		Infer Value of Objects 
4		Infer Maximum Depth Agent Would Jump to Reach Object 

Figure 1: Overview of the 4 tasks in Experiments 1 and 2, including the main manipulations and measures.

**Participants** N=108 adults (48 female, mean age = 39.86 years, range = 23-89 years) were recruited through Amazon Mechanical Turk and were included in the final sample after exclusions. Fourteen participants were excluded for taking less than four minutes 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 from a pilot study. All data collection methods and procedures were approved by the Committee on the Use of Human Subjects at Harvard University. The sample size, participant inclusion criteria, methods, and data analysis plan for this study were formally pre-registered on the Open Science Framework. All data, code, materials, and pre-registration documents can be found at <https://osf.io/u8b9s/>.

**Materials and Design** This four-part experiment was deployed on Qualtrics, an online survey platform. Our stimuli were adapted from events shown to infants in previous work from our lab (Liu et al., under revision). Our stimuli featured an agent (a red smiling sphere), a set of reward objects of varying color and shape (e.g. cones, cylinders, and prisms, etc), and 7 trenches of constant width (2 units in Blender space) but varying depth (1 to 7 units in Blender space). We fixed trench width to control for the physical cost associated with jumping across the gaps. 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. Task 4 always appeared last due to the concern that the explicit language in the experiment (“the deepest trench [the agent] would jump”) could influence participants’ judgments in the other tasks. Participants never saw the agent fall, but they were asked to consider the agent’s mental state in the hypothetical situation if they were to fall in Task 1 (which was randomly assigned to appear first, second, or third across participants).

**Procedure** Following consent procedures, participants were introduced to an agent who interacted with different gaps to reach objects on the other side. Next, they completed comprehension questions on discriminating the depth difference between trenches (i.e. judging which is deeper), and on establishing the ambiguity of the agent’s dispositions towards the cliffs (“Before she acts, do we know which cliffs [the agent] wants to jump?”) and preferences for the objects (“Before she acts, do we know which things she likes?”). Before each of the 4 tasks, participants answered three comprehension 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?”)<sup>2</sup>. If participants answered incorrectly, they were prompted to re-read the question and try again.

**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 ac-

<sup>2</sup>These comprehension questions were designed for child participants, but both children and adults underwent the same procedure.

count for the maximum random effects structure (Barr, Levy, Scheepers, & Tily, 2013), allowing each participant a random slope and a random intercept, we pared down some models to only include a random intercept for participants due to model convergence issues<sup>3</sup>. Our significance threshold was a two-tailed alpha level of 0.05.

We also conducted two exploratory analyses, asking (1) whether linear or non-linear models better accounted for the data in Tasks 1–4 and (2) whether combinations of objective and subjective predictors of the agent’s emotions from Task 1 better accounted for the data in Task 2 (predicting action) and Task 3 (inferring reward). We used Akaike Information Criteria (AICs) as a measure of model fit and parsimony (Sakamoto, Ishiguro, & Kitagawa, 1986). In order to compare statistical models in a robust way, we used bootstrapping techniques to assess the difference in AICs for both exploratory analyses. For more details, see results.

### **Task 1: Do people associate deeper trenches with more negative reward?**

In Task 1, we ask the basic question of whether participants ascribe negative utility to the trenches used in previous studies with infants, and whether their ratings systematically varied with the depth of these trenches.

**Methods** Participants in this task saw images of the agent facing trenches of varying depth (1 unit to 7 units in Blender space) in random order. Participants then rated (1) how the agent would feel as it was jumping and (2) if it fell in, using a scale that ranges from “really unhappy” to “neutral” to “really happy”. The left-right anchors of the scales were counterbalanced between participants and consistent within participants. See Figure 1.

**Results** We found that as the trenches became deeper, people judged that the agent felt more unhappy both as it was jumping, 95% CI [-7.299, -4.765], unstandardized  $B=-6.032$ , standardized  $\beta=-0.474$ ,  $SE=0.644$ ,  $p<0.001$ , and if it were to fall in, [-5.784, -4.910],  $B=-5.347$ ,  $\beta=-0.561$ ,  $SE=0.223$ ,  $p<0.001$ . These two ratings were correlated to each other, [0.445, 0.552],  $r(751) = .500$ ,  $p<0.001$ , 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 ( $M=52.064$ ,  $SD=25.455$ ), [-29.991, -26.707],  $t(752) = -33.898$ ,  $p<0.001$ . This finding shows that participants indeed expected the agent to place more negative reward on deeper trenches, both in terms of its mental states as it faced this obstacle and the alternative state of the world if its actions failed. See Figure 2A.

We explored whether the relationship between trench depth and people’s ratings was linear or nonlinear (here, following exponential decay,  $DV\_Subj\_Jump \sim e^{-IV\_Objective} + (1|ID)$ ) by using bootstrapping to generate difference in

AICs between the linear and non-linear models (sample  $N = 108$  participants with replacement, 1000 iterations)<sup>4</sup>. Across bootstrapped samples, we found that the linear model provided the best balance between fit and parsimony in predicting the agent’s feelings while jumping, bootstrapped 95% confidence interval of the differences between linear and non-linear AICs from 1000 samples (CI) [48.163, 142.046], mean AIC difference across all bootstrapped samples ( $M$ )= 91.361,  $SE=24.350$ , systematic difference between bootstrap distribution and samples (Bias)=7.849 and if it fell, [114.651, 242.602],  $M=173.053$ ,  $SE=33.273$ , Bias=12.690.

### **Task 2: Do people use relative peril to predict others’ actions?**

In Task 2, we ask whether people appreciate that peril can influence others’ future actions by measuring if they expect others to minimize danger, holding equal the physical cost of all possible actions and the rewards these actions lead to.

**Methods** On each trial, participants saw an agent face a choice between jumping 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), generating a depth difference ranging 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 within participants. Each scenario featured a different pair of identical goal objects.

Across 7 trials, participants used a sliding scale to indicate which direction they think the agent will jump, ranging from “definitely left” to “definitely right”, with “either direction” as the midpoint.

**Results** We found that as relative danger between trenches increased, people were more likely to judge that the agent would jump the shallower trench, [10.719, 13.577],  $B=12.148$ ,  $\beta=0.776$ ,  $SE=0.732$ ,  $p<0.001$ . These results suggest 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 Figure 2B.

Next, we explored whether the relationship between relative trench depth and people’s predictions was linear or nonlinear. We bootstrapped (sample of 108, 1000 iterations) differences in AICs between a linear model and logistic ( $DV\_Direction \sim 1/(1+e^{-IV\_Obj\_Depth\_Diff}) + (1|ID)$ ), and found that the logistic model outperformed the linear model in predicting adults’ expectations over the agent’s future action, [-98.139, -29.941],  $M=-59.007$ ,  $SE=17.963$ , Bias=-2.645.

<sup>3</sup>This happened for the hypothesis-driven analysis for Tasks 1, 3, and 4 (adults), and Task 3 (children), and for the exploratory analysis for Tasks 2-3 (adults).

<sup>4</sup>For all models across tasks in both exploratory analyses, we included random intercepts for participants but excluded random slopes due to convergence issues.

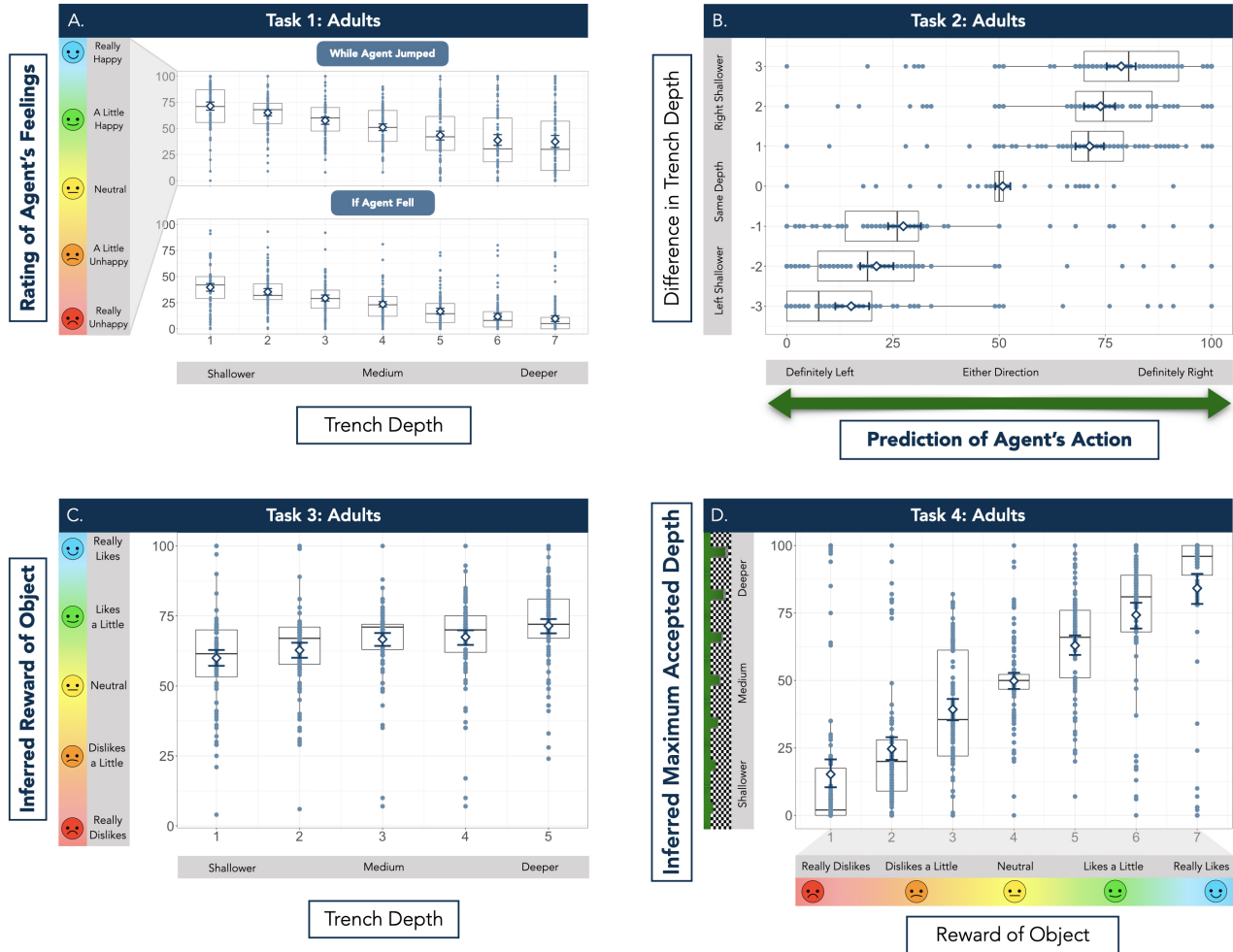


Figure 2: Results from Tasks 1-4 in Experiment 1 (N=108). Bold axis names indicate dependent measures. Individual points indicate raw data, diamonds and error bars indicate means and bootstrapped 95% confidence intervals around the mean, and boxes indicate middle two quartiles of data. (A) People rated that the agent would feel worse while jumping over (top) and if they fell into (bottom) deeper trenches. (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 rating 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 inference for how deep a trench the agent would jump for an object tracked with how much the agent reported liking that object.

To explore how people’s ratings from Task 1 factored into these judgments, we computed the AIC values for seven combinations of three predictor variables: people’s ratings of the agent’s emotions while jumping, their ratings of the agent’s emotions if it fell, and the objective values of cliff depth. Even though all three variables individually predicted people’s judgments ( $p_{\text{objective}} < 0.001$ ,  $p_{\text{fall}} < 0.001$ ,  $p_{\text{jump}} < 0.001$ ), we did not find any model that was better able to balance fit and parsimony compared to the rest<sup>5</sup>.

### Task 3: Do people infer that more perilous actions indicate higher rewards?

In Task 3, we investigate whether people infer the rewards of goals from the peril that others were willing to withstand for those goals, and whether their inferences about value vary quantitatively with manipulations of peril.

**Methods** In Task 3, the main manipulation was how deep of a trench the agent was willing to jump for a goal object. On 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

<sup>5</sup>For details see Gjata (undergraduate thesis, 2020), here: <https://osf.io/9ue6E/>

ranged from "really like" to "really dislike", with "neutral" as the midpoint. Which anchor appeared on the left versus right was counterbalanced between participants and consistent within participants. Trials were shown in random order.

**Results** We found that people's rating for how much the agent liked the goal object varied with how deep a trench the agent previously jumped for that goal object. People's ratings increased as trench depth increased, [2.122, 3.400],  $B=2.761$ ,  $\beta=0.266$ ,  $SE=0.326$ ,  $p<0.001$ . This finding suggests that people use the amount of peril others overcome for their goals to infer the reward associated with these goals. See Figure 2C.

By comparing the difference in AICs between a linear model and a logarithmic model ( $DV_{Preference} \sim \ln(IV_{Depth_{Accept}}) + (1|ID)$ ) across 1000 bootstrapped samples, we found that neither model reliably outperformed the other, [-7.081, 8.992],  $M=-1.022$ ,  $SE=4.035$ ,  $Bias=0.113$ .

We used model AICs to explore whether these judgments were best predicted by the depth of the trench, people's ratings from Task 1, or some combination of the three. Even though all three variables individually predicted people's judgments ( $p_{objective}<0.001$ ,  $p_{fall}<0.001$ ,  $p_{jump}<0.001$ ), we did not find any model that was better able to balance fit and parsimony compared to the rest.

#### **Task 4: Do people infer that others are more willing to withstand higher peril for more valuable goals?**

In Task 4, we ask 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?

**Methods** In Task 4, we manipulated how much the agent reported liking a new set of objects, indicated on the same scale from Task 3. Across 7 trials, the agent reported valuing the object at 7 uniformly spaced levels (from "really dislikes" to "really likes"), and then were asked to rate on a sliding scale the deepest trench the agent would be willing to jump for the object, given how much she likes it. See Figure 1. The left-right anchors of the preference scale was consistent within participants across tasks 3 and 4 but counterbalanced across participants. Trials were presented in random order.

**Results** We found that people's judgments about trench depth varied depending on how highly the agent valued these goal objects. As the reported value of the objects increased, people judged that the agent would jump deeper trenches to reach them, [10.945, 12.609],  $B=11.777$ ,  $\beta=0.710$ ,  $SE=0.424$ ,  $p<0.001$ . This finding suggests that people use other people's known values over goals to infer the amount of peril they would be willing to withstand to obtain these goals. See Figure 2D.

In comparing the differences in AIC values between a linear and logarithmic model ( $DV_{Depth} \sim \ln(IV_{Preference}) + (1|ID)$ ) over 1000 bootstrapped samples, we found that the linear model better fit the data while minimizing model complexity, [21.071, 115.616],  $M=56.826$ ,  $SE=24.049$ ,  $Bias=6.934$ .

## **Experiment 2: Children (in progress)**

Here, we report results from a pre-registered study conducted on 6- to 8-year-old children. Data collection is ongoing, so we refrain from making conclusions on these preliminary results.

### **Participants**

$N=20$  children (12 female, mean age = 83.28 months, range = 72.24–92.03 months) were recruited at the Harvard Lab for Developmental Studies and included in the reported analyses. Our final sample size will consist of 36 participants. We chose to focus on 6- and 7-year-old children based on their ability to successfully use continuous scale measures when reporting their responses (Gweon & Asaba, 2018). For data-sets, data analysis files, and pre-registration see <https://osf.io/u8b9s/>.

### **Methods and Results**

Children saw the same survey as adults from Experiment 1, presented on a tablet by an experimenter. Children inputted all of their own scale responses directly onto the tablet.

**Task 1 Results** Children's predictions for how the agent felt also varied with cliff depth, both as the agent was jumping, [-8.802, -5.202],  $B=-7.002$ ,  $\beta=-.549$ ,  $SE=0.897$ ,  $p<0.001$  and if it fell in, [-7.886, -4.071],  $B=-5.979$ ,  $\beta=-.502$ ,  $SE=0.949$ ,  $p<0.001$ . These ratings were correlated, [0.324, 0.586],  $r(138) = 0.465$ ,  $p<0.001$ , but as with adults, children rated the agent would feel worse if it fell ( $M=23.957$ ,  $SD=23.892$ ), compared to while it was jumping ( $M=54.243$ ,  $SD=25.612$ ), [-34.572, -26.000],  $t(139) = -13.971$ ,  $p<0.001$ .

**Task 2 Results** Children's predictions for where the agent would go depended on the depth difference between trenches, [10.928, 14.639],  $B=12.784$ ,  $\beta=0.754$ ,  $SE=0.947$ ,  $p<0.001$

**Task 3 Results** Children's ratings for how much the agent valued the goal object did not vary with the depth of the cliff the agent jumped for that object, [-4.224, 0.254],  $B=-1.985$ ,  $\beta=-.122$ ,  $SE=1.128$ ,  $p=0.085$ .

**Task 4 Results** Children's ratings for the deepest cliff the agent would be willing to jump for an object scaled with how much the agent reported liking the object, [5.582, 9.940],  $B=7.761$ ,  $\beta=0.463$ ,  $SE=1.108$ ,  $p<0.001$ .

## **Discussion**

When things go wrong, how wrong could they go? Across a 4-task experiment in adults (full study) and in children (data collection in progress), we showed that people are sensitive to how badly actions can end (Task 1), expected others to minimize these negative states (Task 2), and expected others to trade off these negative possibilities against the reward of successfully reaching the intended goal (Tasks 3-4). Altogether, these findings show that people are tuned into the degree of peril and reward that others face, and use these variables to explain and predict others' actions.

These results support and extend the framework theory that

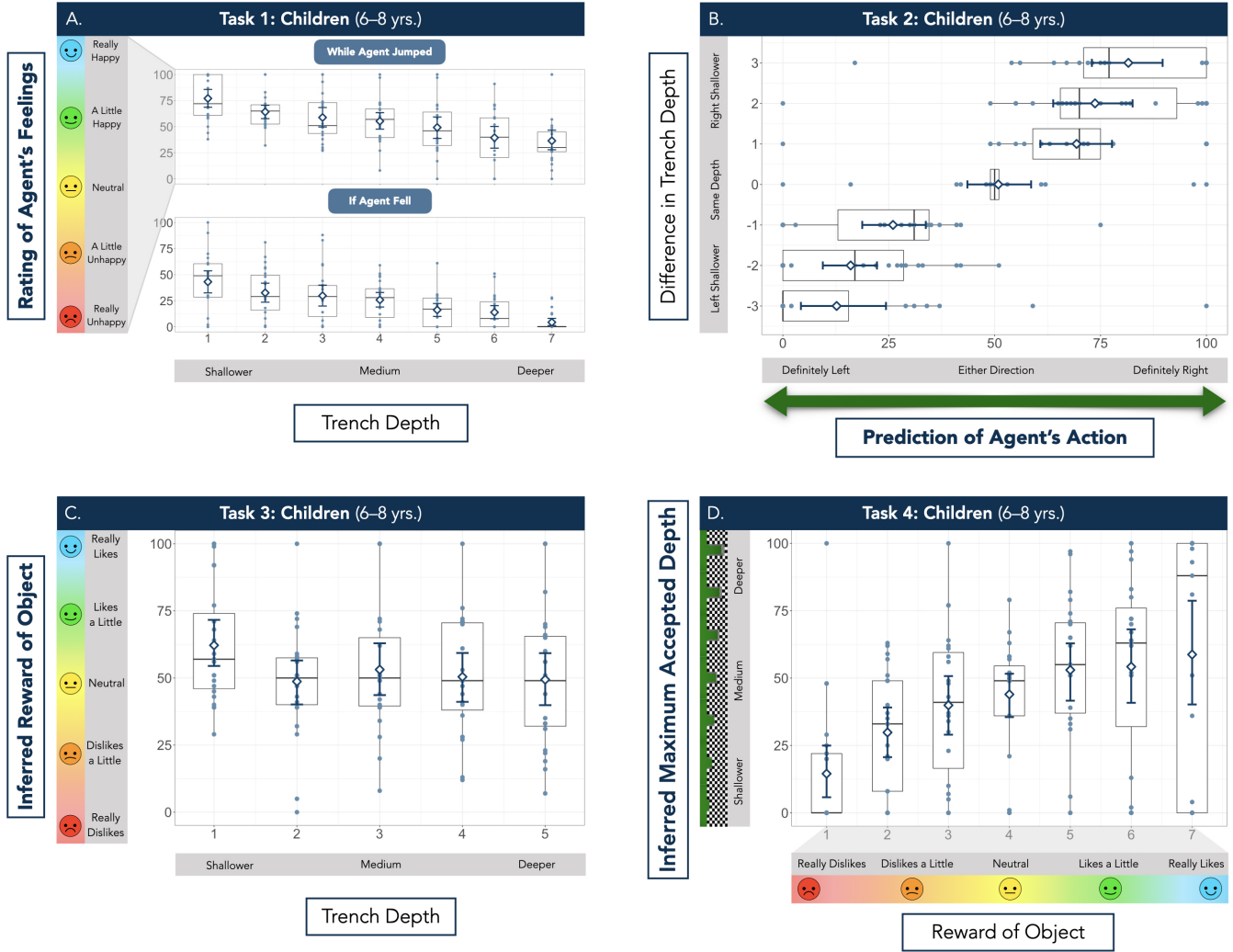


Figure 3: Results from Experiment 2 (N=20 6- to 8-year-old children, data collection in progress). Bold axis names indicate dependent measures. Individual points indicate raw data, diamonds and error bars indicate means and bootstrapped 95% confidence intervals around the mean, and boxes indicate middle two quartiles of data.

we understand other people’s minds and actions by inverting their plans (Bayesian Theory of Mind; Baker et al., 2017, 2009), which can be decomposed into variables like cost and reward (the Naive Utility Calculus; Jara-Ettinger et al., 2016). First, our results suggest 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<sup>6</sup>.

Our data are consistent with at least two possible conceptions of danger. First, people could represent danger,  $D(A)$ ,

<sup>6</sup>Indeed, all the actions that participants viewed involved identical action trajectories, and the stimuli from each task featured just images from these trajectories.

as a kind of negative reward, that, like physical cost,  $C(A)$ , trades off against 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, danger is defined over actions (e.g. jumping over a deep trench), with no explicit representation of the dangerous state itself.

Second, people could 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 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,  $C(A)$ :

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

This second conception of danger explicitly relies on counterfactual representations of possible futures, and the probabilities that these futures will become real. For now, it remains an open question which of these two models is a better description of people's conception of danger.

Our results also leave open the question of what psychological and physical knowledge supports judgments of peril, which could include the agent's emotional states, beliefs about what is possible, and bodily pain as the result of physical injury. While the results from Task 1 indicate that participants can appreciate the mental states taking place during action and the negative bodily consequences of the agent getting hurt, these results do not address how these representations factor into people's judgments of danger.

Here, we focus on peril in a very specific scenario. The perils we and others face in real life are far more complex and entangled with many other factors, like the probability of succeeding or failing (Kahneman & Tversky, 1979; Wellman, Kushnir, Xu, & Brink, 2016), the actor's emotional states and abilities (Skerry & Spelke, 2014; Jara-Ettinger et al., 2015), beliefs about their own utilities (Jara-Ettinger, Floyd, Tenenbaum, & Schulz, 2017), and our own preferences about risk as observers (Liu, McCoy, & Ullman, 2019). Moreover, this work demonstrates that children and adults can be prompted to reason over continuous representations of reward and danger, but do not yet do not show that people spontaneously engage in this type of quantitative reasoning. While this work does not yet capture the richness and complexity of our intuitive action understanding, it provides a test bed for studying these issues. We see the current study as the first step towards a research program that investigates how our minds make sense of the complex social world, and how we grow into this knowledge over development.

## Acknowledgments

We thank our participants, and the Harvard Lab for Developmental Studies for project feedback and support. This work was supported by the Center for Brains, Minds, and Machines (CBMM), funded by National Science Foundation Science and Technology Center award CCF-1231216, by the Harvard College Research Program (to NG), and by a National Science Foundation Graduate Research Fellowship under grant DGE-1144152 (to SL).

## References

- Adolph, K. E. (2000). Specificity of learning: why infants fall over a veritable cliff. *Psychol. Sci.*, *11*(4), 290–295.
- 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. *J. Mem. Lang.*, *68*(3), 255–278.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, *67*(1).
- Dennett, D. C. (1987). *The intentional stance*. London: The MIT Press.
- Gibson, E. J., & Walk, R. D. (1960). The "visual cliff". *Sci. Am.*
- 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. *Am. J. Psychol.*, *57*(2), 243–259.
- Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *J. Exp. Psychol. Gen.*
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.*, *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.
- 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 Dev.*, *84*(1), 226–240.
- Liu, S., McCoy, J., & Ullman, T. D. (2019). People's perception of others' risk preferences. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.
- Liu, S., Ullman, T., Tenenbaum, J., & Spelke, E. (under revision). Dangerous ground: Thirteen-month-old infants are sensitive to peril in other people's actions.
- 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.
- Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81.
- Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216.
- 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.
- Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668–676.