Appendix A
Supplemental Results

Exploratory Analyses

**Task 1 (Adults)**

**Shape of responses:** 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 the difference in AICs between the linear and non-linear models (sample N = 108 participants with replacement, 1000 iterations). 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 1 (Children)**

**Shape of responses:** We explored whether children’s responses in Task 1a were better fit by a linear model or an exponential decay model (\(DV_{Subj\_Jump} \sim e^{-IV\_Objective} + (1|ID)\)) by calculating the difference in AIC values across 1000 bootstrapped samples (sample N = 36 participants with replacement), but found that neither model outperformed the other, \([-5.632, 44.785]\), M = 17.533, SE = 13.248, Bias = 1.911. When performing the same analysis with an exponential decay model (\(DV_{Subj\_Fall} \sim e^{-IV\_Objective} + (1|ID)\)) for Task 1b, we found that the linear model better balanced fit and model complexity, \([8.673, 61.526]\), M = 32.532, SE = 13.513, Bias = 3.012. In other words, we found that an increase in trench depth tracked predicted a corresponding linear increase in the agent’s negative emotions.

---

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

7 We implemented an exponential decay function, (decreasing trend), in Task 1 and a logarithmic function, (increasing trend), in Tasks 3 and 4 for both adults and children following past research that indicates how linear increases in a physical stimulus result in smaller and smaller changes in psychobiological perception (Fechner, 1966). Thus, it is plausible that constant increases in our manipulated variable would lead to decreasing changes in participant’s ratings of the agent’s emotions (Task 1), goal valuation (Task 3), and maximum withstood danger (Task 4).
**Task 2 (Adults)**

**Shape of responses:** 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\(^8\) (DV\(_{Direction} \sim \frac{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. Descriptively, while adults thought the agent could go either way when the two trenches were equally deep (M=51.4, full scale 0-100, 95% CI [44.0, 58.7]) as soon as the trenches differed in depth by Blender 1 unit, adults predicted that the agent would jump the shallower cliff (e.g. for difference of -1, M=24.0 [17.3, 30.7]) and further increases in this depth difference generated proportionately smaller changes in people’s judgments (e.g. for difference of -2, M=16.0 [10.5, 21.5]).

**Height of fall vs appraisal of fall, details:** To explore how people’s inferences about the agent’s appraisals over these obstacles (from Task 1) factored into their predictions in Task 2, above and beyond the main task manipulation of trench depth, we added people’s ratings of the agent’s emotions while jumping (Task 1a) and if it were to fall in (Task 1b) as predictors of their judgments, in addition to the objective depths of the trenches. Above and beyond how deep the trenches were, people’s judgments of how the agent would feel while they were jumping trenches of a particular size from Task 1, predicted their responses when answering questions about the same trenches in Task 2, [-0.189, -0.020], B=-0.105, SE=0.043, \(\beta=-0.073, t(749)=-2.409, p=.0162\). We found a similar and marginal effect for people’s judgments of how the agent would feel if they were to fall, [-0.197, 0.011], B=-0.093, SE=0.053, \(\beta=-0.055, t(749)=-1.755, p=.0797\). In other words, we found that in addition to the physical manipulation of trench depth, for a given pair of trenches, 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. However, we note that these effects are small: adding predictors from Task 1 only added 0.75% variance explained in people’s predictions, relative to our manipulation of the trench depth alone.

\(^8\)We chose the logistic function (S-shaped curve) to implement a plausible non-linear model for Task 2 as it describes the following pattern of results: When the trenches are the same depth, people are uncertain about which way the agent will jump since both reward and effort cost are also kept constant. However, once there is a relative difference in cliff depth, even a small one, people demonstrate a large increase in certainty that the agent will choose the smaller cliff in order to minimize the danger associated with its action. That certainty slowly maxes out as the relative difference between the two cliffs increases.
**Task 2 (Children)**

*Shape of responses:* By comparing the difference in AIC values between a linear model and a logistic model (\(DV_{\text{Direction}} \sim 1/(1+e^{-\text{IV}_{\text{Obj_Depth_Diff}}})+(1|\text{ID})\)), we found that the logistic model provided a better fit while accounting for parsimony, [-35.491, -11.055], M=-22.224, SE=6.262, Bias=-1.869. Thus, like adults, children’s predictions scaled non-linearly with differences in danger across the two options. As with adults, we explored whether children’s ratings of the agent’s appraisals over these situations (Task 1a and 1b) predicted their ratings on Task 2. In contrast to adults, controlling for the main task manipulation, children’s ratings for how the agent felt about trenches before jumping and if it fell did not reliably predict their judgments (Task 1a: [-0.0960, 0.205], B=0.055, \(\beta=0.040, \text{SE}=0.772, t(247)=0.709, p=0.479\)); Task 1b: [-0.046, 0.256], B=0.105, \(\beta=0.075, \text{SE}=0.774, t(247)=1.353, p=0.177\).

*Height of fall vs appraisal of fall, details:* We calculated model AICs to determine whether children’s predictions on the agent’s future actions were best predicted by trench depth, their Task 1 ratings of the agent’s feelings, or some combination of these three factors. Only objective trench depth, not kids’ responses on the agent’s feelings while jumping and if it fell into each cliff, individually predicted their judgements on which cliff the agent would choose to jump in Task 2. The model AICs are as follows: AIC\(_{\text{objective}}=2298.8\), people’s Task 1 judgments of the agent’s feelings about falling into the cliff (AIC\(_{\text{fall}}=2464.7\)) or jumping over it (AIC\(_{\text{jump}}=2460.3\)), by a combination of two variables (AIC\(_{\text{fall_objective}}=2297.7, \text{AIC}_{\text{jump_objective}}=2299.0, \text{AIC}_{\text{fall_jump}}=2454.6\)), or by a combination of all three (AIC\(_{\text{full}}=2299.1\)).

When comparing the models with the two lowest AIC values (1) the model based on only objective depth and (2) the model based on objective depth and the agent’s emotions if it fell, neither model better predicted children’s Task 2 responses, [-15.184, 5.490], M=0.378, SE=5.967, Bias=-1.908.

**Task 3 (Adults)**

*Shape of responses:* Our exploratory analyses of the relationship between people’s judgments and trench depth found both linear and logarithmic models provided an equally reliable fit to the data. By comparing the difference in AICs between a linear model and a logarithmic model (\(DV_{\text{Preference}} \sim \ln(\text{IV}_{\text{Depth_Accept}})+(1|\text{ID})\)) 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.
**Height of fall vs appraisal of fall, details:** To explore how people’s inferences about the agent’s appraisals (from Task 1) factored into their inferences about value from danger (Task 3), we added people’s ratings of the agent’s emotions while jumping (Task 1a) and if it were to fall in (Task 1b) as predictors of their judgments, in addition to the objective depths of the trenches. Above and beyond how deep the trenches were, neither of these ratings for a trench of a particular depth predicted people’s responses (while jumping, 95% CI [-0.083, 0.046], B=-0.018, SE=0.033, \( \beta = -0.027, t(532.890) = -0.553, p = 0.580 \); if fell, 95% CI [-0.031, 0.130], B=0.050, SE=0.041, \( \beta = 0.061, t(529.710) = 1.204, p = 0.229 \)).

**Task 3 (Children)**

In Task 3, participants gauged how much the agent valued different objects based on which cliffs it was willing to jump. While children’s responses did not differ significantly from the null model, we report exploratory analyses here. By comparing the difference in AIC values 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 outperformed the other, [-4.465, 8.992], \( M = -1.872, SE = 1.133, Bias = -0.148 \).

As with Task 2, we compared model AICs to explore whether children’s judgments were best predicted by trench depth, ratings from Task 1a and 1b, or a combination of these three predictors. None of these three variables individually predicted children’s judgments (\( p_{objective} = 0.096, p_{jump} = 0.22, p_{fall} = 0.19 \)). AIC values for each model are as follows: AIC\(_{objective} = 1589.8\), Task 1 judgments of the agent’s feelings about falling into the cliff (AIC\(_{fall} = 1590.9\)) or jumping over it (AIC\(_{jump} = 1591.1\)), by a combination of two variables (AIC\(_{fall\_objective} = 1591.6\), AIC\(_{jump\_objective} = 1591.7\), AIC\(_{fall\_jump} = 1592.5\)), or by a combination of all three (AIC\(_{full} = 1593.5\)). We took the two models with the lowest AIC values, (1) the model incorporating objective depth and (2) the model incorporating the agent’s emotions if it fell into various cliffs, and calculated the difference in their AIC values. Across 1000 bootstrapped samples, we found that there was no significant difference in their performance, [-0.366, 13.474], \( M = 6.354, SE = 3.518, Bias = -0.073 \).

**Task 4 (Adults)**

**Shape of responses:** In comparing the differences in AIC values between a linear and logarithmic model (\( DV_{Depth} \sim \ln(DV_{Preference}) + (1|ID) \)) over 1000 bootstrapped samples, we found that the linear model provided a better fit the data while minimizing model complexity, [21.071, 115.616], \( M = 56.826, SE = 24.049, Bias = 6.934 \).
**Task 4 (Children)**

**Shape of responses:** Across 1000 bootstrapped samples, we found no significant difference between the parsimony and fit of a linear model versus a logarithmic model (\(DV_{\text{Depth}} \sim \ln(IV_{\text{Preference}}) + (1|ID))\), [-9.560, 8.069], \(M=-1.793\), \(SE=4.417\), Bias=0.453.

**Comparing children and adults**

Did children respond to information about depth (Tasks 1-3) and value (Task 4) differently to adults? To ask this question, we compared responses from adults and children in all tasks by adding an interaction between group (children vs adults) and our primary manipulation (In Tasks 1-3, trench depth, and Task 4, object value).

We found that the depth manipulation produced a marginally stronger effect in Task 1b (how would the agent feel if she were to fall in?) in children than adults, \(\beta=-0.142\), 95% CI [-2.970, 0.039], \(B=-1.466\), \(SE=0.768\), \(t(142.036)=-1.909\), \(p=.058\).

We also found a stronger effect in adults than children in Task 3 (inferring reward from danger), \(\beta=-0.334\), 95% CI [-5.753, -2.781], \(B=-4.267\), \(SE=0.758\), \(t(574)=-5.628\), \(p<.001\), and a stronger effect in adults than children in Task 4 (predicting danger from reward), \(\beta=0.287\), 95% CI [3.026, 6.595], \(B=4.810\), \(SE=0.911\), \(t(861.35)=-5.283\), \(p<.001\).

In Task 1a and Task 2, children’s and adults’ responses to the manipulations of depth were not significantly different from each other.

**Appendix B**

**Supplemental Figures**

For all of the following figures: bold axis names indicate dependent measures, connected individual points indicate data from a single participant, diamonds and error bars indicate means and bootstrapped 95% confidence intervals around the mean, and boxes indicate middle two quartiles of data.
Results from pilot experiments (N=54 adults recruited through Amazon Mechanical Turk). Results are compiled across three different versions (V1, V2, and V3), but the most significant changes across versions were made in Task 3. In Versions 1 and 3 of Task 3, participants saw the agent jump one cliff and refuse another cliff one unit deeper across trials. In Version 2, participants only saw the deepest cliff that the agent would jump, which matched with the accepted cliff each trial in Versions 1 and 3. Bold axis names indicate dependent measures. Individual points indicate raw data, and lines connect data from a single participant. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals around the mean, and boxes indicate middle two quartiles of data.
Figure B2

Results from pilot experiments conducted prior to main results (N=36 children, 18 female, mean age = 7.04 years, range = 6.03–8.58 years). Results are compiled across four different versions (V1, V2, V3, and V4), but the most significant changes across versions were made in Task 3. As with adults, in Versions 1 and 3 of Task 3, participants saw the agent jump one cliff but refuse a deeper cliff, while Version 2, participants only saw the deepest cliff that the agent would jump. Version 4 recruited only 4 participants, whose responses are counted in Task 1, Task 2, and Task 4, but omitted from Task 3 since the experimental design differed greatly from Versions 1-3 of that task. Bold axis names indicate dependent measures. Individual points indicate raw data, and lines connect data from a single participant. Diamonds and error bars indicate means and bootstrapped 95% confidence intervals around the mean, and boxes indicate middle two quartiles of data.
Figure B3

Results from pilot experiments that explore the null result for children in Task 3 (N=16 children, mean age = 6.70 years, range = 6.12–7.89 years), across four versions. Each child saw two versions. First, they completed either Version 1 (N=5), Version 2 (N=5), or Version 4 (N=6). Next, all children completed Version 3 (N=16). In Versions 1 and 2, children were shown what an agent would do for one object. Either they saw the agent jump one cliff, but refuse a cliff two units deeper for the same object (Version 1) or they saw the deepest cliff the agent would jump for a toy (Version 2). In Versions 3 and 4, children saw what the agent would do for two different objects and were asked which object the agent preferred. As with Versions 1 and 2, in Versions 3 and 4 respectively, children either saw the agent accept a cliff and refuse another to reach each toy, or they only saw the deepest cliff the agent would jump to get to each. While objects changed in each trial, we kept constant what the agent would do for one toy (i.e. it would only jump the shallowest depth), but varied across trials the cliffs which it would jump for the other toy.
Figure B4

Results from Task 1a in adults, where responses from each participant is plotted in each grid (N=108 total).
Figure B5

Results from Task 1b in adults, where responses from each participant is plotted in each grid (N=108 total).
Figure B6

Results from Task 2 in adults, where responses from each participant is plotted in each grid (N=108 total).
Figure B7
Results from Task 3 in adults, where responses from each participant is plotted in each grid (N=108 total).
Figure B8

Results from Task 4 in adults, where responses from each participant is plotted in each grid (N=108 total).
Figure B9

Results from Task 1a in children, where responses from each participant is plotted in each grid (N=36 total).
Figure B10
Results from Task 1b in children, where responses from each participant is plotted in each grid (N=36 total).
Figure B11

Results from Task 2 in children, broken down by responses from each single participant (N=36 total).
Figure B12

Results from Task 3 in children, where responses from each participant is plotted in each grid (N=36 total).
Figure B13
Results from Task 4 in children, where responses from each participant is plotted in each grid (N=36 total).
Figure B14

Results from each Task 1a and 1b in adults (A,C) and children (B,D), broken down by whether participants saw this task 1st, 2nd, or 3rd.
Figure B15
Results from Tasks 2 and 3 in adults (A,C) and children (B,D), broken down by whether participants saw this task 1st, 2nd, or 3rd.
Figure B16

Results from a new sample of adult participants (N=34) who completed Tasks 3 and 4 in counterbalanced order, broken down by whether participants saw Task 3 first or Task 4 first.