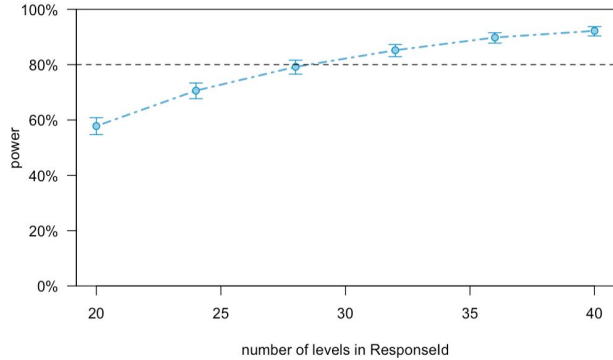
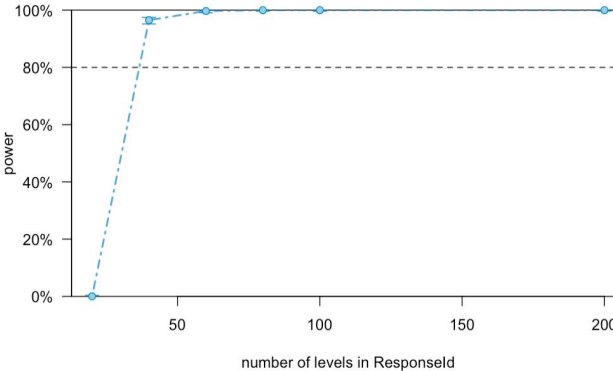
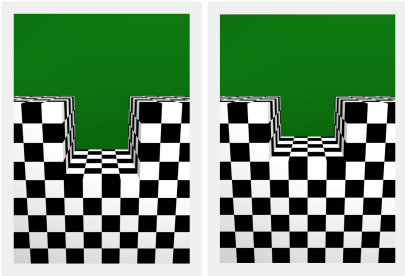


This pre-registration document has been approved by all researchers involved in the project as of 12/20/19.

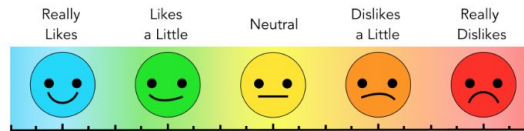
Title	Look Before You Leap: Reasoning About the Danger and Rewards in Other People's Actions																												
Start Date	Dec 20, 2019																												
Researchers	Removed for anonymity																												
Experimenters																													
Sample Size	<p>N = 36 children and 108 adults, after exclusions</p> <p>We performed a power analysis on the mixed-effects model for both the kid and adult pilot data of Part 2 (which showed the weakest effect of the three tasks - see below) to determine the amount of participants needed to detect an effect with power 0.8 and alpha 0.05. The power analysis indicated that we need ~ 30 kids and ~ 50 adults to reach this threshold. Therefore, we decided it would be feasible to collect data for 36 kids. Given that MTURK data collection is more efficient than in-lab experiments and precise quantitative effects of our manipulations would be useful for future modeling studies, we set the sample size to 104 adults. Graphs are shown below:</p> <p style="text-align: center;">Kid Experiment (Part 2)</p>  <table border="1"> <caption>Kid Experiment (Part 2) Power Analysis Data</caption> <thead> <tr> <th>number of levels in Responselid</th> <th>power</th> </tr> </thead> <tbody> <tr><td>20</td><td>58%</td></tr> <tr><td>25</td><td>72%</td></tr> <tr><td>30</td><td>80%</td></tr> <tr><td>35</td><td>88%</td></tr> <tr><td>40</td><td>92%</td></tr> </tbody> </table> <p style="text-align: center;">Adult Experiment (Part 2)</p>  <table border="1"> <caption>Adult Experiment (Part 2) Power Analysis Data</caption> <thead> <tr> <th>number of levels in Responselid</th> <th>power</th> </tr> </thead> <tbody> <tr><td>0</td><td>0%</td></tr> <tr><td>25</td><td>95%</td></tr> <tr><td>50</td><td>100%</td></tr> <tr><td>75</td><td>100%</td></tr> <tr><td>100</td><td>100%</td></tr> <tr><td>150</td><td>100%</td></tr> <tr><td>200</td><td>100%</td></tr> </tbody> </table>	number of levels in Responselid	power	20	58%	25	72%	30	80%	35	88%	40	92%	number of levels in Responselid	power	0	0%	25	95%	50	100%	75	100%	100	100%	150	100%	200	100%
number of levels in Responselid	power																												
20	58%																												
25	72%																												
30	80%																												
35	88%																												
40	92%																												
number of levels in Responselid	power																												
0	0%																												
25	95%																												
50	100%																												
75	100%																												
100	100%																												
150	100%																												
200	100%																												

Participant Age	Children 6-8 years old, US Adults 18 years of age or older recruited on Amazon Mechanical Turk
Methods and Design	<p>In this project, we ask (1) whether people expect others to quantitatively trade off the danger associated with goal-directed actions against how rewarding the goals are. As a second, exploratory question, we are interested in (2) whether those tradeoffs are better predicted by the physical properties or people's subjective assessments of dangerous obstacles. To test the first question, we created 3 different tasks. We start by testing whether (a) people appreciate that danger can influence others' future actions by measuring if they expect others to minimize danger all else being equal. Then, we test the tradeoffs more directly, by assessing whether (b) people infer the rewards of goals from the danger that others were willing to incur for those goals and (c) people expect others to incur more danger for more highly-valued goals. To test the second question, we ask whether people's tradeoffs between danger and reward are best predicted by objective properties of the physical environment (i.e. how deep different cliffs are) or by people's subjective perceptions of how an agent would feel if they fell into different cliffs.</p> <p>Children will come into the Lab for Developmental Studies to access a touch-screen tablet. Adults will participate online through Amazon Mechanical Turk, a platform that crowd-sources the recruitment process. Following consent procedures, participants are introduced to the goal of this experiment: to infer an agent's preferences and predict its future actions.</p> <p>Throughout the experiment, participants will see multiple trials featuring different objects, some of which the agent likes more than others, and different cliffs, some of which are deeper than others. There are four distinct tasks in this experiment. Before each part, we explain the setup and ask comprehension questions relevant to the scale measure used in that part.</p> <p>Part 1: Given equal reward, predict action</p> <p>We show participants two cliffs that have the same object on either side. One cliff remains fixed at a medium depth while the other varies from shallow to deep. Across 7 trials, participants answer which direction they think the agent will jump and how certain they are on a continuous sliding scale that ranges from "definitely left" to "definitely right".</p> <p>Part 2: Given danger, predict value</p> <p>In the second part, we show participants cliffs of varying depth and indicate which ones the agent is willing to jump. Across 5 trials, the agent accepts one cliff and refuses to jump another that is two units deeper. Then, we ask how much they like the object on a continuous sliding scale that ranges from "really like" to "really dislike."</p> <p>Part 3: Given value, predict danger</p> <p>In the third part, we show participants the same preference scale from Part 2, but vary the slider's position, indicating how much the agent likes an object. Across 7 trials, we ask participants about the deepest cliff the agent would</p>

	<p>jump to reach that object on a continuous sliding scale from “very shallow” to “very deep.”</p> <p>Part 4: Subjective danger</p> <p>In the fourth part, we show participants an agent faced with cliffs of varying depth, and measure the amount of danger they associate with those cliffs by asking (1) how the agent would feel as it was jumping and (2) if it fell in, using a scale that ranges from “really happy” to “really unhappy.”</p> <p>If this subjective measure is correlated with objective cliff depth, then we can test the second exploratory question on which is a better predictor of people’s inferences about reward and predictions about what an agent will do.</p> <p>Counterbalances and Randomization</p> <p>Across participants, we will counterbalance tasks 1, 2, and 4 using a Latin Square, but leave Part 3 last due to potentially leading wording: asking about “the deepest cliff” that an agent would jump indicates how reward should influence depth more strongly than the other experiments. Within the experiments, we counterbalance whether the left or right cliff varies in depth in part 1 and the left-right anchors of the scales in part 2, 3 and 4. Anchors for preference and happiness scales are consistent within participants; for example, if a participant is assigned to a positive-negative valenced version, scales will consistently range from “like” to “dislike” and “happy” to “unhappy”. All participants will see one set of objects per task with no overlap in color-object combinations between tasks. Within tasks, each object set is arranged in one of two order configurations across trials in order to counterbalance object and cliff depth/preference associations. Finally, within a task, the trial order is randomized for each participant.</p>
Stimuli	<p>This four-part experiment was designed on qualtrics, an online survey platform. The scenarios in each part, including agent, objects, and cliffs, were designed using Blender. Scales were designed using Keynote.</p> <p>Comprehension Check Examples:</p> <p>Baseline Comprehension Check Example</p> <p>Which cliff is deeper?</p> 

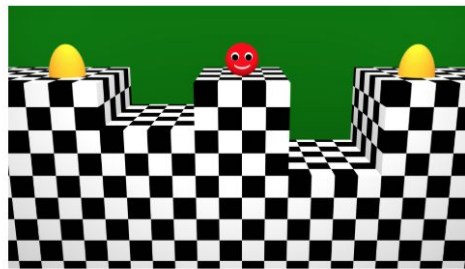
Scale Comprehension Check Example

Where would you put the slider if you think Wendi **really dislikes** the thing?



Experiment 1: Inferring Future Action Given Danger Example

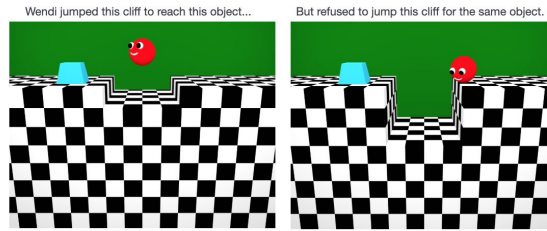
Here's Wendi and here are two objects that she can get.



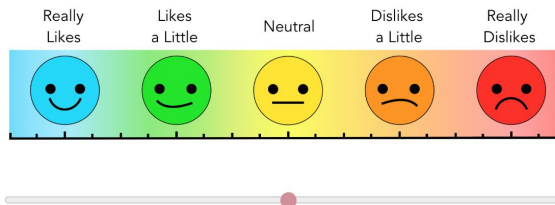
Which way will Wendi jump?



Experiment 2: Inferring Value Given Danger Example



How does Wendi feel about this object?

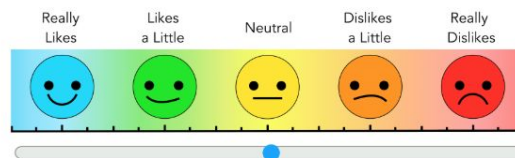


Experiment 3: Inferring Danger Given Value Example

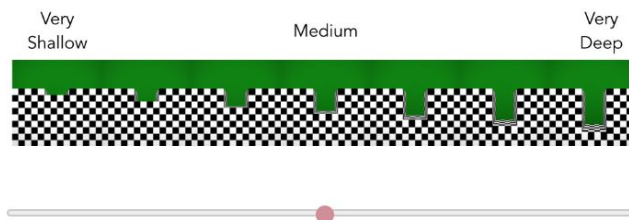
We asked Wendi about this object:



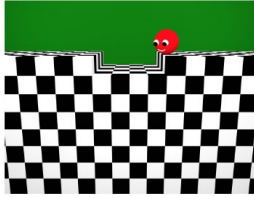


This is how Wendi felt about this object:



What is the deepest cliff that Wendi would be willing to jump for this object?



Experiment 4: Subjective Measure Example

	<p>Imagine that Wendi tried to jump over this cliff:</p>  <p>How would Wendi feel while she was jumping?</p> <div data-bbox="670 520 1167 642"> <p>Really Happy A Little Happy Neutral A Little Unhappy Really Unhappy</p>  </div> <p>How would Wendi feel if she fell in?</p> <div data-bbox="670 821 1167 963"> <p>Really Happy A Little Happy Neutral A Little Unhappy Really Unhappy</p>  </div>
<p>Logic and Hypothesis</p>	<p>Previous work suggests that infants consider danger when reasoning about the decision-making processes of other people. For example, if they see an agent jump a deeper cliff for Object A than Object B, then they expect that the agent likes Object A more. In this study, we test children and adults in similar scenarios to analyze whether people quantitatively form these expectations. For example, as a cliff increases in depth between trials, does preference also increase or does it remain fixed? The logic is as follows:</p> <p>(1) If people expect others to plan over continuous representations of danger and reward, and quantitatively trade off amongst them:</p> <ul style="list-style-type: none"> ❖ In Part 1: then given constant reward, when faced with two cliffs that are equal in depth, people will be uncertain about which cliff the agent will jump. As the first cliff becomes deeper, then people should be more certain that the agent will traverse the second cliff for the reward. Conversely, as the first cliff becomes shallower, then people should be more likely to say that the agent will jump the first cliff instead. In other words, as relative danger between cliffs increases, people should be more certain that the agent will jump the shallower cliff thereby minimizing danger with reward and cost held constant. ❖ In Part 2: then as the cliff depth that the agent would jump increases, so should people's rating of how much the agent likes that object. ❖ In Part 3: then as the amount that the agent likes the object increases, so should the depth of the deepest cliff the agent would jump to reach it.

	<p>(2) If subjective perception of danger better explains people's judgments and predictions, then the statistical model with subjective perception should provide a better fit across experiments than the statistical models that incorporate objective cliff depth as determined by units in Blender space.</p>
Primary DV	<p>The hypothesis-driven measures consist of children and adult's responses on the continuous scales in each experiment.</p> <p>Participants will be asked:</p> <p>In Part 1: To predict the agent's action by indicating on the scale whether the agent will go definitely left to either direction to definitely right.</p> <p>In Part 2: How much they think the agent likes a goal object on a scale that ranges from really dislikes to neutral to really likes.</p> <p>In Part 3: How deep a cliff the agent would be willing to jump for an object on a scale that ranges from really shallow to medium to really deep.</p> <p>In Part 4: How an agent would feel as it jumped and if it fell into different cliffs on a scale that ranges from really sad to neutral to really happy.</p>
Primary analysis	<p>Primary Analysis</p> <p>(1) We plan on using a linear mixed effects model (lme4 R package), using random effects to account for within subject variability. Analyses will be performed separately for kids and adults. We will use likelihood ratio tests to assess whether our experimental manipulation (expressed in the hypothesis-driven model) explains more variance in the data than the null model. We will use a two-tailed alpha level of 0.05 as our significance threshold.</p> <p>Analysis for Experiment 1:</p> <pre>#null model null <- lmer(data = exp1, formula = DV_Direction ~ 1 + (1+IV_Obj_Depth_Diff ResponseId) # hypothesis-driven model modell <- lmer(data = exp1, formula = DV_Direction ~ IV_Depth + (1+IV_Obj_Depth_Diff ResponseId) # likelihood ratio test for comparing models anova(null, modell) Analysis for Experiment 2: #null model</pre>

```

null <- lmer(data = exp2, formula = DV_Preference ~ 1 +
(1+IV_Depth_Accept|ResponseId)

# hypothesis-driven model

model2 <- lmer(data = exp2, formula = DV_Preference ~
IV_Depth_Accept + (1+IV_Depth_Accept|ResponseId)

#LRT

anova(null, model2)

```

Analysis for Experiment 3:

```

#null model

null <- lmer(data = exp3, formula = DV_Depth ~ 1 +
(1+IV_Preference|ResponseId)

# hypothesis-driven model

model3 <- lmer(data = exp3, formula = DV_Depth ~
IV_Preference + (1+IV_Preference|ResponseId)

#LRT

anova(null, model3)

# ResponseId = a unique value tracking each subject

```

The code for each experiment generates two models: a null model based on random variability in the data and a hypothesis-driven model that incorporates the independent variable into the null model to predict our dependent measures. The likelihood ratio test compares the two models to assess whether the hypothesis-driven model better explains the data.

Secondary Analysis (Exploratory)

(2) We plan to compare whether people's subjective assessments of danger are correlated with objective cliff depth. Given a correlation, we then plan to model people's responses based on the objective depth of cliffs versus their subjective judgment on the danger associated with those cliffs. In cases where models are nested, we will use LRTs to assess model fit. Otherwise, we will compare model AICs and BICs to assess fit.


```
subj_corr <- lmer(data = exp4, formula= IV_Objective ~  
DV_Subj_Diff_Fall + DV_Subj_Diff_Jump +  
(1+IV_Objective|ResponseId))
```

this code generates the correlation between objective
and subjective values of danger

If a correlation exists:

Comparing Subjective v Objective Danger Assessments (Exp. 1)

```
sub_model_full <- lmer(data = exp1, formula =  
DV_Direction ~ IV_Obj_Depth_Diff + DV_Subj_Diff_Fall +  
DV_Subj_Diff_Jump + (1_Obj_Depth_Diff|ResponseId))
```

```
sub_model_fall <- lmer(data = exp1, formula =  
DV_Direction ~ IV_Obj_Depth_Diff + DV_Subj_Diff_Fall +  
(1_Obj_Depth_Diff|ResponseId))
```

```
sub_model_jump <- lmer(data = exp1, formula =  
DV_Direction ~ IV_Obj_Depth_Diff + DV_Subj_Diff_Jump +  
(1_Obj_Depth_Diff|ResponseId))
```

```
anova(model1, sub_model_full, sub_model_fall,  
sub_model_jump)
```

Comparing Subjective v Objective Danger Assessments (Exp. 2)

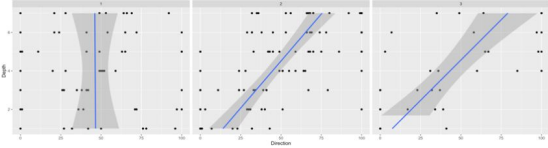
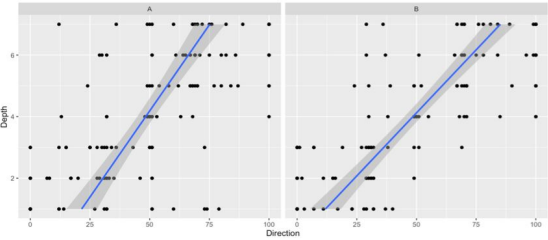
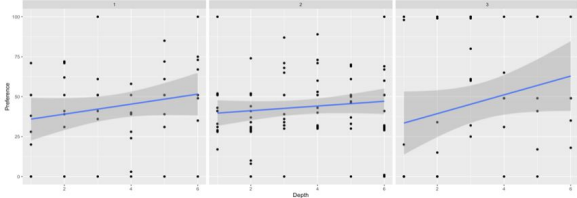
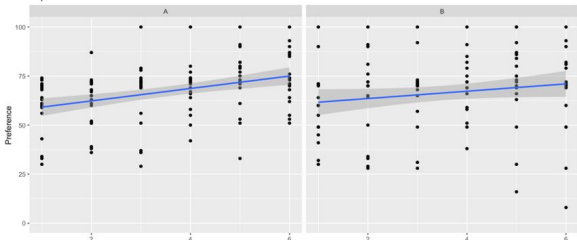
```
sub_model2_full <- lmer(data = exp2, formula =  
DV_Preference ~ IV_Depth_Accept + DV_Subj_Fall +  
DV_Subj_Jump + (1+IV_Depth_Accept|ResponseId))
```

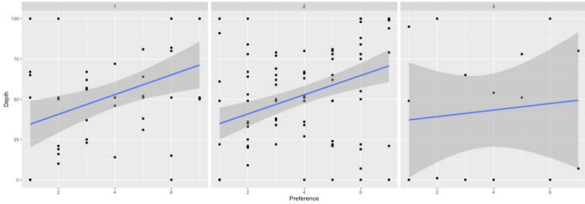
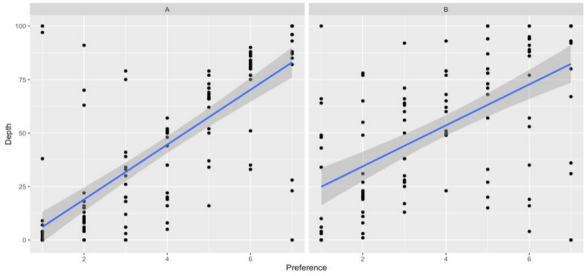
```
sub_model2_fall <- lmer(data = exp2, formula =  
DV_Preference ~ IV_Depth_Accept + DV_Subj_Fall +  
(1+IV_Depth_Accept|ResponseId))
```

```
sub_model2_jump <- lmer(data = exp2, formula =  
DV_Preference ~ IV_Depth_Accept + DV_Subj_Jump +  
(1+IV_Depth_Accept|ResponseId))
```

```
anova(model2, sub_model_full, sub_model2_fall,  
sub_model2_jump)
```

this code compares the original hypothesis driven model to a second
model incorporating participants' subjective rating of danger

	<p>Further exploratory analyses may include examining age and gender effects, or using non-linear models, like logistic or curvilinear functions, to test for best fit.</p>
Pilot Results	<p>36 Kids and 38 Adults were tested in the pilot sample across three versions of the study. Pilot data will not be included in the full sample. Figures are shown below:</p> <p>Experiment 1:</p> <p>Kids Results (Versions 1, 2, and 3)</p>  <p>Adult Results (Versions 1 and 2)</p>  <p>Experiment 2:</p> <p>Kids Results (Versions 1, 2, and 3)</p>  <p>Adult Results (Versions 1 and 2)</p>  <p>Experiment 3:</p>

	<p style="text-align: center;">Kids Results (Versions 1, 2, and 3)</p>  <p style="text-align: center;">Adult Results (Versions 1 and 2)</p> 
Exclusion Criteria	<p>MTURK Specific:</p> <ul style="list-style-type: none"> ❖ Task Exclusion: If < 4 trials are completed in any section, that specific part will be excluded but the rest of that participant's data will be included in analysis ❖ Participant Exclusion: <ul style="list-style-type: none"> ➢ If data for 3 parts are excluded ➢ If time to complete survey < 4 min ➢ If participant fails attention check (if > 5 characters are incorrect when retyping phrase) ❖ For the question, "what is your job in this experiment": participant must include at least 1 keyword from "cliff", "object", "like", "dislike", "jump", "want", "action", "judge", "deep", "shallow", "rate", "small", "big", "answer", "figure"; otherwise all of that participant's data will be excluded. <p>Lab Specific:</p> <ul style="list-style-type: none"> ❖ Task/Trial Exclusion: <ul style="list-style-type: none"> ➢ If experimenter makes a significant error in the wording of one trial, then that trial will be excluded ➢ If experimenter makes a serious error in the introduction of a part, then that section's worth of data will be excluded ➢ If completed trials < 4 in any section, that specific task will be excluded (e.g. if child elects to end participation) ❖ Participant Exclusion: <ul style="list-style-type: none"> ➢ Parental interference (e.g. supplies an answer, speaks to child during at least 1 experimental measure). ➢ Technical failure (computer issues) ➢ Participant is not comfortable with English ➢ If data for 3 parts are excluded

	<p>If our actual sample size after exclusions exceeds our pre-specified participant sample size (N=36 for kids, 108 for adults), then we will include that data in our analysis. We will aim to run exactly the number of participants as pre-specified, but there may be a case where we accidentally run more adult participant on mTurk than indicated above when compensating for exclusions.</p>
Exclusion Decisions	<p>Participants will be excluded from the analysis as the experiment is running if they meet any of the exclusion criteria. Critically, all decisions to exclude data will be made using the above objective criteria <i>without looking at the data</i>.</p>
Other analysis details	<p>Data will be analyzed once we have collected all participants accounted for by the power analysis. We will visualize the data for occasions like lab meetings and research meetings, but we will not run any analyses until data collection is finished.</p>