Modeling Risk in Physical Settings

Linette Kunin (linette@mit.edu)
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology

Catherine Mei (meic1212@mit.edu)
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology

Sabrina Piccolo (piccolo@mit.edu)
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology

Abstract

People make intuitive judgements about the riskiness of actions to determine which action to take (e.g. to get $100 for sure, or to flip a coin for a chance to get $200). However, how people use the perceived likelihood of consequences and the severity of the consequences to calculate this risk in *physical situations* has not been investigated. We asked 45 adults about 9 scenarios in which an animated agent was in the process of crossing a bridge of various widths over a cliff of various depths. In each, we asked the participants to rate the likelihood of failure (the probability that the agent will fall), the severity of the consequences (how bad falling would be), and the risk of the situation as a whole. We modeled these responses and found that the model that best approximates people's intuitions on risk is the geometric mean between an individual's likelihood and severity rating for similar ratings of likelihood and severity, and minimum of the two values when there is a larger disparity between them. This highlights a relationship between likelihood of failure or severity of consequences in shaping risk perception that may be specific to physical situations: when either variable is extremely low or extremely high, only one of them will drive intuitions about risk.

Keywords: risk; modeling; cognitive science; adults

Introduction

Before we cross a street, we make an assessment of how dangerous it is to cross. We might consider factors such as the number of cars on the street, how fast the cars on the street are traveling, how wide the street is, and whether the pedestrian crossing signal is on. This study focuses on the phenomenon of risk assessment in physical domains and situations. Many of our judgments are based on how risky actions are and how much reward we get from them. Generally, we choose risky actions when rewards are greater. Humans also have some intuition about the level of risk associated with a situation. But how is the risk calculated? Behavioral experiments have shown that people use the expected values associated with choices to make decisions and that these expected values are determined by the probability of the values occurring multiplied by the magnitude of the reward (Jara-Ettinger et al., 2020; Von Neumann & Morgenstern, 1944). In this study, we hypothesize that judgments of risk are determined similarly in physical scenarios, where neither magnitude nor probability are directly observable. Here, we look at likelihood of failure and consequences of failure as possible moderators of how risky an action or situation is.

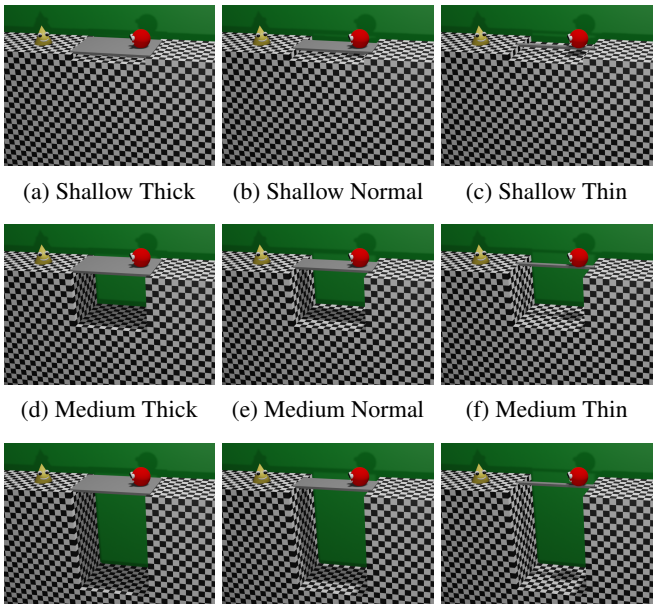
Related Work

At an early age, humans possess an understanding of danger and risk. For example, infants become less likely to cross a bridge over a drop-off that they could fall into when the width of the bridge decreases and tend to avoid deep drop-offs more than shallow drop-offs (Kretch & Adolph, 2013). Furthermore, when infants at least a year old watch an agent leap over cliffes of different depths toward a goal, they not only expect the agent to choose the least dangerous path, but they also determine which goal the agent prefers based on the risk the agent is willing to take to reach it (Liu et al., 2022). Other research has shown that both children and adults use the degree of danger to explain and predict others' actions and they also expect agents to minimize the danger of their actions (Gjata et al., 2022). Considering that people have intuitions about danger and ultimately risk, this raises the question about what factors shape these perceptions of risk. In economics, expected value is calculated by multiplying the amount of a reward by the probability of the reward occurring (Myerson et al., 2011); perhaps calculating risk involves a similar relationship between the perception of danger and the likelihood of danger.

Methods

For this study, we recruited 45 neurotypical adult volunteers from the MIT community and our personal circles, such as family and friends. The stimuli consisted of 9 still-images of a red ball agent attempting to cross a gray bridge propped on top of a gap to reach a yellow cone agent. We used three settings for bridge width: thin, normal, thick. Similarly, we used three settings for the gap height: shallow, medium, deep. The normal width and the medium height were the same distance from the thin and the thick bridge and the shallow and deep cliff respectively. Each bridge width and gap height setting were combined to form the 9 scenarios in our experiment.

These stimuli were inspired by the experimental design of Kretch & Adolph (2013) in which crawling infants had to decide whether to cross a narrow or wide bridge over a deep or shallow drop-off. The stimuli used for this study are presented in Figure 1.



(a) Shallow Thick (b) Shallow Normal (c) Shallow Thin

(d) Medium Thick (e) Medium Normal (f) Medium Thin

(g) Deep Thick (h) Deep Normal (i) Deep Thin

Figure 1: Stimuli Scenarios

For data collection, we created a Google Form with each of the 9 scenarios. For each scenario, we asked the participant to answer 3 questions: How likely is it that the red character will fall? How bad would it be if the red character fell? How risky is the path that the red character is taking? Participants answered each question on a Likert scale from 1 to 7, where the scale ranged from not at all likely to very likely, not at all bad to very bad, and not at all risky to very risky, respectively.

We posited that it is possible for the first scenario in the survey to establish a strong baseline or prior for the participant, which might then impact participant judgments for the remaining stimuli. Therefore, we created 3 different Google Form surveys, each with a different ordering of the stimuli scenarios. We collected 15 responses for each survey, resulting in a total of 45 survey responses.

Results

Do the physical parameters we manipulated actually evoke impressions of likelihood and severity?

We collected data about three metrics in our study. These three metrics were likelihood of falling (corresponding to survey question: How likely is it that the red character will fall?), the severity of the fall (corresponding to question: How bad would it be if the red character fell?), and the risk associated with the scenario (corresponding to survey question: How risky is the path that the red character is taking?).

In order to investigate the distribution of our data, we visualized participants' responses to likelihood and severity across the three different levels of bridge width and cliff depth. Figure 2 shows the distribution of responses to the likelihood and severity question for the 9 scenarios. The data is visualized using a violin plot, where we partitioned the data based on bridge width and cliff depth for both the likelihood and severity questions.

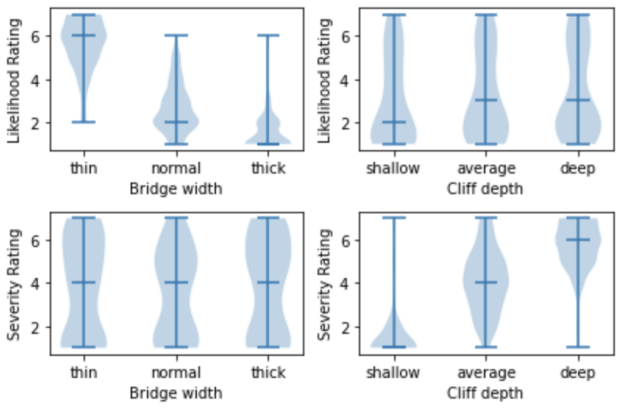


Figure 2: Distribution of Likelihood and Severity Responses. Row 1 is the Likelihood rating distribution. Row 2 is the Severity rating distribution. Column 1 shows the data partitioned on bridge width, and column 2 shows the data partitioned on cliff depth. Middle line is the median and upper and lower bars are the extremes.

If our manipulation worked as intended, then (i) the width of the bridge that the agent crossed should influence people's impressions of the likelihood it will fall, (ii) the depth of the cliff below the bridge should influence people's impressions of how bad it would be if the agent fell, and (iii) these manipulations should only affect the judgment that it is paired with in (i) and (ii). As expected, we find a dependency between the bridge width and likelihood rankings. Specifically, as bridge width increases, the likelihood rating decreases. Additionally, the likelihood ranking is largely unaffected by the depth of the cliff. We can also see that the severity of falling is dependent on the depth of the cliff. Specifically, as cliff depth increases, so does the severity rating. Severity rating is generally unaffected by the width of the bridge.

These are the judgments we were trying to elicit when we created the stimuli and questions. When we designed our experiment, we expected to be able use the physical parameters cliff depth and bridge width as proxies for judgments about severity and likelihood of falling. When we varied the physical parameters in each scenario, we were successful in eliciting a wide range of human judgments about likelihood, severity, and risk of falling, which then allowed us to model the relationship between these three values.

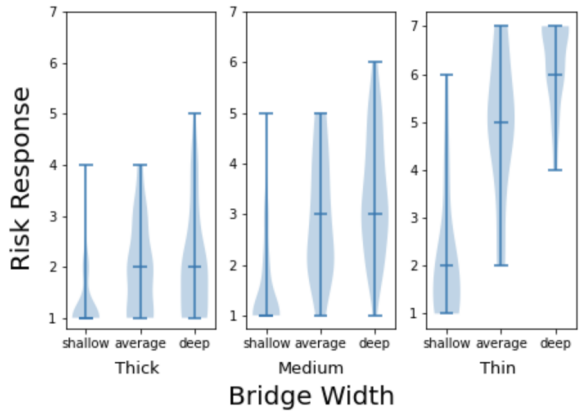


Figure 3: Distribution of Risk responses as function of bridge width and cliff depth. Data is first grouped by bridge width and then split by cliff depth. Middle line is the median and upper and lower bars are the extremes.

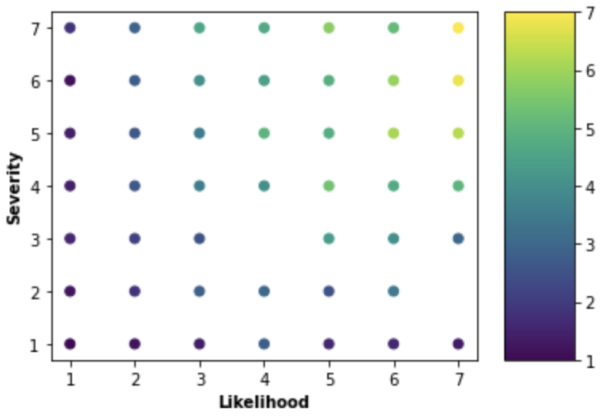


Figure 4: Risk vs Likelihood and Severity. Response to risk on the Likert Scale is shown in color.

Are judgments of physical risk influenced by manipulations over bridge width and cliff depth?

In order to investigate how the distribution of risk is correlated with the physical parameters (cliff depth and bridge width) in our experiment, we created a violin plot showing the data first grouped by bridge width, then split by cliff depth within each bridge width bucket. Figure 3 shows the resulting distribution.

From the violin plot, we can see that peoples' judgments about risk vary with the bridge width and cliff depth, but that the effect of bridge width varies by cliff depth. In particular, deeper cliffs generally lead to higher assessments of risk when the width of the bridge is held constant. On the other hand, thinner bridges generally lead to higher assessments of risk when the cliff depth is held constant. From the data, we also observed that for shallow cliff depths, the assessment of risk was almost always lower than for average or deep cliffs, regardless of the bridge width. However, when the bridge was really thick, the assessment of risk was almost always lower than the cases where the bridge was normal width or thin. These series of observations could suggest that judgments about risk are not merely a simple function of the product between these two manipulations.

Do judgments of physical risk relate to likelihood and severity judgments?

We hypothesized that risk assessment for a particular stimuli scenario would be dependent on both the likelihood and severity of the agent falling. Before attempting to model risk as a function of these judgments, we wanted to visualize our data and evaluate the plausibility of our hypothesis.

To visualize possible relationships between risk and the distribution of severity and likelihood, we created the following scatterplot. We grouped our participant responses to all 9 scenarios based on their ranking of likelihood and severity of falling. It was possible that multiple responses mapped to the same (likelihood, severity) pair. To generate a more representative risk value for a particular pair, we averaged all of the risk judgments for that (likelihood, severity) pair. This average risk is shown using a colored dot in Figure 4.

As expected, we notice that as people's judgments of likelihood and severity of falling increase, so do their judgments about the risk. Additionally, we noticed that for very small values of likelihood of falling, the risk was small regardless of severity, while for very small values of severity of falling, the risk was small regardless of likelihood.

These observations align with our intuition and hypothesis. In particular, if the possibility of falling is very low, then the risk of injury will be very low even if the consequences of falling are severe since the agent is not likely to fall. Similarly, if the consequences of falling are not severe, even if the agent does fall, the situation is still not risky. This again suggests that our intuitive understanding of risk cannot be encapsulated by one model. Our goal is to take these intuitive observations and create a representative computational model of human assessment of risk.

What generative model over likelihood and severity best describes people's physical risk judgments?

Our goal was to model the relationship between likelihood, severity, and risk. We determined 6 possible models that take input parameters likelihood and severity to estimate output predicted risk. These models are listed below:

(1) **Likelihood:** Risk = likelihood

(2) **Severity:** Risk = severity

(3) **Mean:** Risk = mean(likelihood, severity)

(4) **Minimum:** Risk = min(likelihood, severity)

(5) **Maximum:** Risk = max(likelihood, severity)

(6) **Geometric Mean:** Risk = $(\text{likelihood} \cdot \text{severity})^{\frac{1}{2}}$

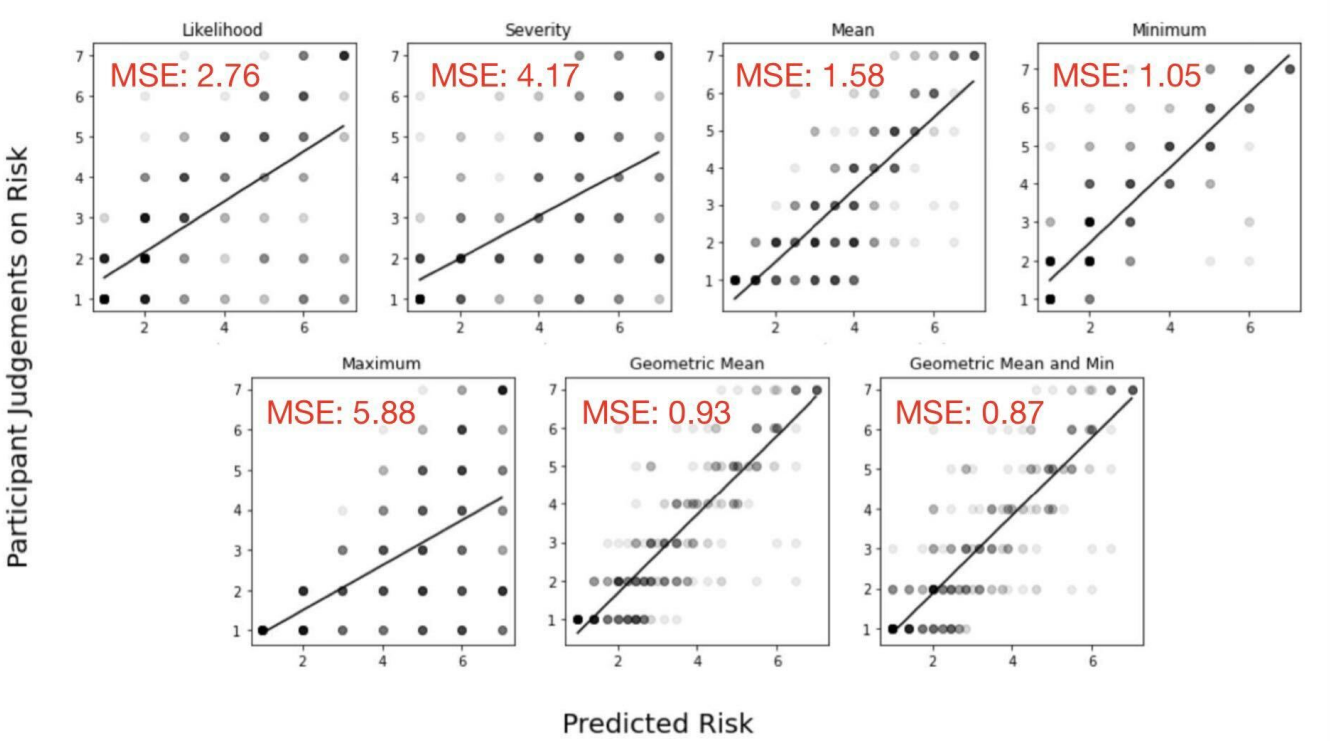


Figure 5: Model predicted risk plotted against human risk judgments for each participant response using each of the 7 models. Lines of best fit are shown for each model.

The first model always equates risk with the likelihood judgment, meaning that the model would assume risk is directly determined by likelihood. The second model always equates risk with the severity judgment, meaning that the model would assume risk is directly determined by severity. In the third model, risk is calculated as the average of the likelihood and severity judgments. Given the likelihood and severity judgments, the fourth and fifth models equate the risk with the minimum and maximum of the two options, respectively. Finally, the sixth model equates risk with the geometric mean of likelihood and severity.

Because our participants have different backgrounds and experiences, we expected large variability in risk judgments between different participants. Therefore, we decided to apply our model to each participant's judgment separately. Essentially, we took a single participant's response to the likelihood and severity questions for a certain scenario, applied one of the 6 models, and produced a predicted risk. We then compared the predicted risk to the actual risk rating, across all trials, across all participants. These values are plotted in Figure 5.

To evaluate the performance of our models, we used the following procedure. For each scenario, we calculated the mean squared error between the predicted risk and the actual risk rating. Essentially, we took the difference between the predicted risk and actual risk rating, squared the difference, and averaged these squared differences across all data points. Once we calculated the mean squared error for each function and scenario, we averaged the mean squared error for each function across the nine scenarios to find the average mean squared error per function. Table 1 summarizes our finding for each model.

Model	MSE
Maximum	5.88
Severity	4.17
Likelihood	2.76
Mean	1.58
Minimum	1.05
Geometric mean	0.93
Geometric mean and Min	0.87

Table 1: Average Mean Squared Errors for each model, calculated by averaging the mean squared error per scenario across subjects for each model. Models are listed from worst to best performance.

As mentioned previously, human intuition about risk is complex, and so we believed that a single calculation could not capture this complex relationship entirely. To refine our approach, we decided to combine the two models that had the best performance: the geometric mean model (MSE = 0.93) and minimum model (MSE = 1.05).

Recall that one of the key conclusions we reached when analyzing the correspondence between severity, likelihood, and risk was that when judgments about severity or likelihood of falling were very small, the risk would be small regardless of other parameters. To create a composite model, we first took

an average of all of the likelihood ratings and severity ratings per question. We used these two metrics to estimate the distribution of responses per question. As an example, if the average likelihood for a certain scenario was large, then we hypothesized that most likelihood ratings that scenario were high.

In our revised model, we applied the minimum function to our data when the average likelihood and average severity ratings differed by a large amount, which we defined as having a difference of at least 4. When the average likelihood and average severity ratings differed by less than 4, we applied the geometric mean function to our data. We felt that it was sensible to apply the minimum function to our data when the average ratings differed by a lot because we observed that extreme low ratings in either likelihood or severity tended to dominate over other factors when determining risk. When the two averages were comparable, we found that the geometric mean produced the most representative model of human judgments of risk.

Comparing the performances of the models in Table 1 shows that a combination of geometric mean for smaller differences in likelihood and severity, and minimum for larger differences in likelihood and severity has the smallest mean squared error (MSE = 0.87). This implies that out of the above models, the combined geometric mean and minimum model, flexibly applied when the worst case scenario was both moderately bad and moderately likely (geometric mean) versus when the worst case scenario was either very bad but very unlikely, or very likely but not very bad (minimum), best approximates people's empirical judgements on physical risk in these scenarios.

Finally, we created visualizations of our results to evaluate differences between the models. For each of the 7 models, we plotted the predicted risk against the participant judgment of risk across the 9 scenarios. We then generated a line of best fit for each of the scatterplots. The results are shown in Figure 5. For models that predicted the actual risk judgments by humans well, we expected the line of best fit to be most similar to a 45° line (i.e. the line actual judgments = predicted risk). From qualitative observations of the lines of best fit for the 7 models, we see that the geometric mean model and the composed geometric mean and minimum model reflected human risk judgments the most.

Discussion

The results of this study reveal that the ways human perceptions of likelihood of failure and severity of consequences interact and ultimately affect human perceptions of risk is not a simple function of expected value. Our best performing model for reflecting human intuitions about risk in regards to our stimuli is a combination of using geometric mean for smaller differences in likelihood and severity and using minimum for larger differences in likelihood and severity. This reveals that when likelihood of failure or severity of consequences is either extremely low or extremely high, only one of the two will drive intuitions about risk, but this is not the case when the two are similar in value. In other words, this study not only supports the intuition that likelihood of failure and severity of consequences play a role in risk calculations, but it also reveals ways that these two variables interact to shape the perception of risk.

Limitations and Future Directions

Although our study provides insight into human perceptions of risk as factors of likelihood of failure and severity of consequences, it is important to note some limitations of the work. For example, this study uses simple scenarios with only two variables, which could have elicited simplified judgements that never actually occur in the real world. Also, the wording of our questions and the fact that we asked each participant to rate the likelihood and severity of each scenario before rating risk could have artificially driven our participants to use their ratings of likelihood and severity to make their judgements on risk. In future studies, a more optimal experimental design may be to include each stimuli scenario image 3 times, each time with a question on one of likelihood, severity, and risk. Then, to eliminate bias from question ordering, all of the 27 individual questions would be scrambled.

Another limitation arose from commentary from our participants. A few participants in our study mentioned that it was unclear if the agent could fall off the entire checkered surface instead of just the bridge and that this could influence ratings in terms of severity and, consequently, risk. In a future experiment, we could design the stimuli so that it is clear that the agent can not fall off the board but only into the canyon. It is also important to note that we created only three surveys but considering each survey could have started in nine different ways, which makes it unlikely that any effect we found is driven entirely by trial order, we did not fully counterbalance the stimuli.

Participants' personal experiences and prior knowledge may have interfered with their judgments as well. For example, one participant commented that the agent crossing the bridge being round could impact judgments because someone may assume that the agent is more likely to roll.

Also, because many of our participants have experience with computational models, they may have been more predisposed to use likelihood and severity to estimate risk in a similar way, skewing the results in favor of our models. Lastly, we didn't have enough participants to split our data into training and testing sets, so we used all of our data to create, test, and then choose the models in this study. In a future study, we could test an independent set of participants' judgements on likelihood, severity, and risk to make sure that in the process of creating our model, we didn't overfit our data.

Author Contributions

The following details each group member's contribution to the project. Please note that most of the project was completed together in synchronous meetings.

Linette Kunin: contributed to designing stimulus and participant survey; contributed to data collection; created code for data analysis and visualization; contributed to final paper.

Catherine Mei: contributed to designing participant survey; contributed to data collection; worked on code necessary to process and model survey data; worked together with team members on data analysis and paper.

Sabrina Piccolo: contributed to designing participant survey; contributed to data collection; provided information about relevant literature; provided feedback on code; wrote sections of the paper.

Acknowledgments

We would like to thank Dr. Shari Liu for providing feedback and guidance. We would also like to thank Saxelab and our peers, friends, and families for participating in our studies. Finally, we would like to thank Dr. Josh Tenenbaum and the course TAs for their support throughout the semester and for designing a course that provided us with the resources and opportunity to conduct this study.

Code and Data

The code and data can be found here: Github Link

References

Gjata, N. N., Ullman, T. D., Spelke, E. S., & Liu, S. (2022). What could go wrong: Adults and children calibrate predictions and explanations of others' actions based on relative reward and danger. *Cognitive Science*, *46*. Retrieved from https://doi.org/10.1111/cogs.13163 doi: 10.1111/cogs.13163

Jara-Ettinger, J., Schulz, L., & Tenenbaum, J. (2020). The naïve utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*. Retrieved from https://doi.org/10.1016/j.cogpsych.2020.101334 doi: 10.1016/j.cogpsych.2020.101334

Kretch, K. S., & Adolph, K. E. (2013). No bridge too high: Infants decide whether to cross based on the probability of falling not the severity of the potential fall. *Developmental Science*, *16*, 336-351. Retrieved from https://doi.org/10.1111/desc.12045 doi: 10.1111/desc.12045

Liu, S., Pepe, B., Kumar, M. G., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2022). Dangerous ground: One-year-old infants are sensitive to peril in other agents' action plans. *OpenMind*, *6*, 211-231. Retrieved from https://doi.org/10.1162/opmi_a_00063 doi: 10.1162/opmi_a_00063

Myerson, J., Green, L., & Morris, J. (2013). Modeling the effect of reward amount on probability discounting. *Journal of the experimental analysis of behavior*, *95*. Retrieved from https://doi.org/10.1901/jeab.2011.95-175 doi: 10.1901/jeab.2011.95-175

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.