Cognition 265 (2025) 106216

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Brief article

People make suboptimal decisions about existential risks☆

Adam Elga [a], Jian-Qiao Zhu [b], Thomas L. Griffiths [b,c,*]

[a] Department of Philosophy, Princeton University, United States of America
[b] Department of Computer Science, Princeton University, United States of America
[c] Department of Psychology, Princeton University, United States of America

ARTICLE INFO

Keywords:
Existential risk
Rationality
Heuristics
Resource allocation

ABSTRACT

Allocating resources to maximize the probability that humanity survives a set of existential risks has a different structure from many decision problems, as the objective is the product of the probabilities of desired outcomes rather than the sum. We derive the optimal solution to this problem and use this solution to evaluate the choices that people make when presented with decisions that have this multiplicative structure. Our participants (total N=2,072) are appropriately sensitive to how responsive a risk is to investment, but are conservative in their decisions and do not allocate enough resources to risks with lower probability of survival. This pattern persists even with alternative framings that emphasize survival probabilities. Our results highlight a systematic flaw in people's intuitions about how to respond to existential risks, and suggest that people may have particular difficulty with decisions that involve multiplicative objectives.

1. Introduction

There has been increasing interest (Bengio et al., 2023; Bostrom, 2013; Greaves & MacAskill, 2021; MacAskill, 2023; Ord, 2020; Slovic, 2020) in reducing the risk of catastrophes that would lead to "the premature extinction of Earth-originating intelligent life" (Bostrom, 2013), such as deadly pandemics, bioweapons (Millett & Snyder-Beattie, 2017), asteroid strikes (Voosen, 2023), or even runaway artificial intelligence (Carlsmith, 2023; Critch & Russell, 2023). Thinking about how to avoid extinction by mitigating these risks involves a distinctive kind of decision problem, as our incentives are *multiplicative*: we primarily care about whether *any* of the catastrophes occur. That is because existential risks threaten consequences so extreme that if one of them is realized, it matters much less which of the others are realized as well. This contrasts with the *additive* effects of risks more commonly addressed in classical decision theory, where if one risk is realized it still can greatly matter which others are also realized (Savage, 1954; Von Neumann & Morgenstern, 1947). Reasoning about such risks is an important special case of reasoning about nonlinear systems. Since decision-makers find reasoning about nonlinear systems difficult (Dutt, 2013; Mckenzie & Liersch, 2011; Olsson et al., 2006; Van Dooren et al., 2003), it is worth asking how effectively they assess plans for minimizing existential risk.

In this paper, we analyze problems that have this multiplicative structure to determine how resources should be optimally allocated to reduce overall existential risk, and show that human decision-makers systematically deviate from this optimal solution. Complementing previous work on the evaluation of compound probabilities in general (Bar-Hillel, 1973; Fan et al., 2019; Nilsson et al., 2013; Wedell & Moro, 2008) and on judgments of the badness of existential risks in particular (Schubert et al., 2019), our results provide a guide to how to best engage with existential risks and a warning about how our intuitions might lead us to fail to do so. They also confirm that people find decisions with multiplicative objectives unintuitive, suggesting that other cases where such decisions arise may also be problematic.

Societies contemplating existential risks have to decide how to allocate their resources to best mitigate those risks. Should we, for example, divide our resources roughly evenly between various disaster-prevention projects? Or should we "put all of our eggs in one basket" and focus on just the most probable cause of extinction? To address these questions, consider the following problem (see Fig. 1). A decision-maker must divide a fixed budget among projects to mitigate several independent existential risks. For each individual risk i, the increase in the probability that we survive that risk is proportional to the amount v_i that we allocate to it. Each risk has a different degree of addressability a_i, as some risks are more responsive to investment than others, and a different baseline probability of survival s_i, reflecting differences in the current chances of catastrophes. The decision-maker's goal is to choose

☆ This work and related results were made possible with the support of the NOMIS Foundation. We thank Ben Enke, Sanjeev Kulkarni, Danny Oppenheimer, and Pietro Ortoleva for helpful discussions. A. Elga thanks Mindy and Gene Stein for research environment support.

* Correspondence to: 320 Peretsman Scully Hall, Princeton University, Princeton NJ 08540, United States of America.
E-mail address: tomg@princeton.edu (T.L. Griffiths).

https://doi.org/10.1016/j.cognition.2025.106216
Received 22 September 2024; Received in revised form 31 May 2025; Accepted 8 June 2025
Available online 21 July 2025

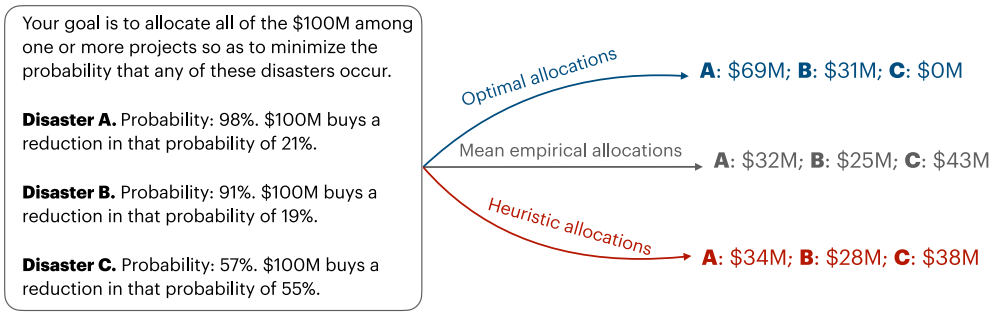

Fig. 1. An example allocation task shows the deviations in mean empirical allocations from the optimal allocations. The probability of a disaster occurring is $1 - s_i$, while the reduction in that probability by investing the entire budget of $100M for that disaster is a_i.

allocations v_i to maximize the probability that we survive all of the risks, $\prod_i (s_i + a_i \cdot v_i)$.

In this setting, a decision-maker should be sensitive to both the baseline survival probabilities and the addressability of risks. Intuitively, if s_i (the baseline survival probability of risk i) is high we should not be too worried about it, and if a_i (the addressability of risk i) is low there is nothing we can do about it. In fact, we show below that what matters to the optimal solution is the *ratio* of these two factors, s_i/a_i. The optimal strategy prioritizes investing in risks based on this ratio. This entails that all else equal, a risk deserves more investment when it can be mitigated more cheaply (a_i is large). No surprise there. But it also entails that all else equal, a risk deserves more investment when our chance of surviving it (s_i) is lower (Cotton-Barratt et al., 2020; Thorstad, 2023). For example if we face catastrophes with survival probabilities of 4% and 40%, it is more valuable to increase the probability of surviving the first catastrophe from 4% to 9% than it is to increase the probability of surviving the second catastrophe from 40% to 80% (since we care about the product of the survival probabilities and $9\% \cdot 40\% > 4\% \cdot 80\%$). This dependence of optimal allocations on baseline chances of survival is a distinctive and somewhat counterintuitive feature of situations in which incentives are multiplicative (Lewis et al., 2023; Lewis & Simmons, 2020).

To understand how people allocate resources in conditions of multiplicative uncertainty, we placed participants in the role of executive director of an organization dedicated to reducing existential risk (see Fig. 1). Given a description of baseline chances and degrees of addressability for three existential risks, participants divided a fixed risk-mitigation budget among the risks with the goal of minimizing the probability that any of the associated catastrophes occur.

2. Model 1: Optimizing for human survival

In our formulation, a decision-maker must allocate a total budget T to mitigate n independent existential risks to maximize overall survival probability. For each individual risk i, the increase in probability that we survive that risk is proportional to the amount v_i that we allocate to it, so that the probability of surviving risk i is given by

$$s_i(v_i) = s_i(0) + a_i \cdot v_i, \tag{1}$$

where each risk i is assumed to have a fixed degree of addressability $a_i > 0$ and baseline probability of survival $s_i(0) > 0$. For simplicity we sometimes write $s_i = s_i(0)$, but here we express $s_i(v_i)$ as a function to clarify the optimization problem. We assume that the budget is not large enough to completely mitigate any individual risk: $s_i(T) < 1$ for each i.

An allocation $\boldsymbol{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ is *optimal* exactly if it maximizes the overall survival probability

$$S(\boldsymbol{v}) = s_1(v_1) \cdot s_2(v_2) \cdots s_n(v_n) \tag{2}$$

subject to the budget constraints that each $v_i \geq 0$, and $v_1 + v_2 + \cdots + v_n = T$. For each risk i we define a *complacency* $c_i(v_i) = s_i(v_i)/a_i$. We write $c_i = c_i(0)$ for the complacency of risk i in the status quo, and note that $c_i(v_i) = c_i + v_i$. Without loss of generality, assume that the risks are ordered so that $c_1 \leq c_2 \leq \cdots \leq c_n$. We write $\overline{c}_m$ for $\Sigma_{i=1}^{m} c_i/m$, the mean complacency of the first m risks.

In the supplement (see Appendix A) we prove (Proposition 1) that the unique optimal allocation is the allocation $\boldsymbol{v}$ for which

$$v_i = \begin{cases} T/k^* + (\overline{c}_{k^*} - c_i) & \text{if } i \leq k^*, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where k^* is the greatest $k \leq n$ for which $T/k + (\overline{c}_k - c_k) > 0$. Risks with lower complacency should thus, if they are sufficiently concerning, receive greater investment: the degree to which the complacency differs from the mean complacency determines the degree to which the optimal allocation differs from an equal allocation.

3. Experiment 1: Assessing people's decisions about existential risks

Experiment 1 investigated whether people's decisions about existential risks align with the optimal solution derived in the previous section. We preregistered this experiment at osf.io/t9z3v.

3.1. Methods

3.1.1. Participants

We recruited a total of 1199 participants through the Prolific Academics platform (`www.prolific.com`). Among these, 783 individuals (348 males, 268 females, 167 who chose not to disclose their gender) successfully completed the 15-min experiment and passed the catch trial. The participant age range spanned from 19 to 85, with a median age of 38. Participants were required to be from the US and have a track record of at least 100 submissions with an acceptance rate of 95% or higher. Participants received a base monetary compensation of $3.00 (hourly rate of $12.00). The experimental sessions were carried out in Nov 2023. Informed consent was obtained from all participants. (Princeton University IRB number 10859: "Computational Cognitive Science").

3.1.2. Stimuli

We initially created a grid comprising 120 unique allocations. Each allocation is composed of three non-negative numbers which sum to 1. This structure represents the distribution of resources across three specific mitigation projects.

The generated grid also serves as the optimal solution (shown as blue dots in Fig. 2a, 3a, and 4a) for which we now aim to identify the corresponding tasks. These tasks are characterized by a combination of baseline survival probabilities and their addressability. In each

iteration, we independently generate six numbers within the range of $[0.01, 0.99]$ to represent $s_1, s_2, s_3, a_1, a_2, a_3$ respectively, ensuring that for all i, $s_i + a_i < 1$.

Following this, we calculate the optimal solution for the simulated tasks and identify the grid allocation that best matches it using the absolute differences. The sum of the absolute differences across the three allocations must also be less than 0.02. Should the best-matching grid allocation lack a corresponding simulated task, we directly link it to the current task. In cases where an existing task is already linked to the best-matching grid, we evaluate the degree of correspondence between the optimal solutions of the existing and the newly simulated tasks. The task with the higher level of correspondence is then associated with the grid allocation. This process is repeated until each of the 120 grid allocations is matched with a corresponding task.

3.1.3. Procedure

Before proceeding to the main experiment, all participants were required to successfully complete a comprehension check. This prerequisite ensured that they had a clear understanding of the experimental procedures and objectives. Each trial in the main experiment was structured to assess participants' decision-making processes in the context of three distinct existential risks. Each risk was characterized by two parameters: its addressability and the probability of the corresponding disaster occurring (i.e., $1 - s_i$, where s_i is the baseline probability of surviving that risk).

Participants were tasked with allocating a hypothetical budget of $100 million across these three risks. The objective was to minimize the probability that any of three independent disasters occur. The effect of investing in each disaster was described in terms of how much that investment would reduce the probability of that disaster. Participants were required to articulate the rationale behind their allocation decisions in a provided text box. An example trial is displayed in Figure A5.

For a given set of three pairs of survival probability and degrees of addressability, the optimal model predicts a specific budget allocation among the three corresponding risks. To ensure a broad range of variability in the allocation predicted by the optimal model, we developed a total of 120 unique trials each consisting of 3 randomly generated pairs of experimental parameters with the constraint that fully investing in any given risk does not completely mitigate it. Each participant was randomly assigned to complete 6 of these trials. The order of the 3 mitigation projects was randomized across participants. The catch trial was randomly inserted somewhere between the second trial and the final trial.

3.2. Results

People's allocations systematically departed from optimal allocations in two ways: First, as shown in Fig. 2a, on average people allocated resources more evenly among the risks than was optimal. We used the entropy of a normalized allocation as a measure of its dispersion. The mean entropy of optimal allocations across the 120 tasks was 0.83, while the mean for participant allocations was significantly higher at 1.03, $t(119) = 11.71$ ($p < .01$) (see Fig. 2c). This observed bias towards equitable distributions in mean participant allocations aligns with previous findings on naive diversification strategies (Benartzi & Thaler, 2001; Fox & Rottenstreich, 2003).

Second, while participants (correctly) tended to allocate more resources to risks with higher addressability (a_i) (Fig. 2e), they were not as successful at allocating more resources to risks with lower baseline chances of survival (s_i) (Fig. 2d). We quantified this by investigating the marginal effects of the experimental variables s_i and a_i on allocation decisions. As shown in Fig. 2, the experimental variables were separately categorized into 6 equally sized bins. The outcomes of our statistical tests are compiled in Tables 1 and 2. The findings indicate that the predictions of the optimal model statistically significantly diverge from the human data in all but 3 categories, indicating a poor fit to the human data.

Finally, we explore the possibility that participants misinterpreted the experimental instructions, leading them to maximize the *sum* rather than the *product* of individual survival probabilities. This constitutes an incorrect formulation of the problem. The optimal resource allocation strategy under this additive risk setting is straightforward: concentrate all resources on the risk with the highest addressability (a_i). However, as detailed in the supplement linked in Appendix A, this additively-optimal model fails to accurately capture the patterns observed in the human data.

4. Model 2: A simple heuristic

In an attempt to capture how people allocate resources to mitigate existential risk we defined a simple heuristic model. Intuitively – and in qualitative agreement with the optimal model – people may reason that (i) existential risks with higher baseline survival probabilities require less investment, as survival is already likely, and (ii) existential risks with higher addressability warrant greater investment, as the expected per-dollar returns from mitigating these risks are higher. However, whereas the optimal model explicitly uses the ratio of these two quantities to guide allocation decisions, our heuristic model posits that individuals may instead treat baseline survival probability and addressability as more separate, independent considerations guiding their allocation decisions. Therefore, in this model the investment in risk i is a decreasing function of its baseline survival probability s_i, and an increasing function of its addressability a_i. The model is defined in two steps. First, each risk i is associated with an *investment tendency* w_i:

$$w_i = -\beta_s \frac{s_i}{\sum_{j=1}^{n} s_j} + \beta_a \frac{a_i}{\sum_{j=1}^{n} a_j}, \tag{4}$$

where $\beta_s, \beta_a \geq 0$ are free parameters that adjust the weight given to baseline survival probability and degree of addressability, respectively. The sum in the denominators ensures that the values are normalized across all risks. Second, the allocation to each risk is given by a softmax of the investment tendencies:

$$v_i^{\text{heuristic}} = \frac{e^{w_i}}{\sum_{j=1}^{n} e^{w_j}}. \tag{5}$$

We fitted the heuristic model by minimizing the mean squared errors (MSE) between the model predictions and the individual allocations. This optimization employed the Nelder–Mead algorithm, which evaluates the error at a set of points and extrapolates beyond those points to choose the next point to query (Nelder & Mead, 1965). The heuristic model's best-fitting parameters were $\beta_s = 0.97, \beta_a = 2.39$. The higher β_a value suggests higher participant responsiveness to degrees of addressability than to baseline survival probabilities in decision-making.

In comparison to the optimal model, the heuristic model demonstrated a stronger correlation with mean participant allocations. Specifically, the optimal model achieved a Pearson correlation coefficient of $r = 0.74$ ($p < .01$), whereas the heuristic model that was best-fitted to the data exhibited a Pearson correlation coefficient of $r = 0.84$ ($p < .01$). To account for differences in the number of free parameters between the optimal and heuristic models, we additionally computed the adjusted R^2 as follows: $R^2_{\text{adjusted}} = 1 - (1 - r^2)\frac{n_{\text{datapoints}} - 1}{n_{\text{datapoints}} - n_{\text{parameters}} - 1}$. The resulting adjusted R^2 for the heuristic model is 0.71, which is higher than the optimal model's 0.55. The heuristic model's predictions also showed a closer correspondence with the human data on the statistical criteria used to assess the optimal model above, showing statistically significant differences from the human data in only 4 out of the 12 analyzed categories (see Tables 1 and 2).

To assess how participants diverge from optimal resource allocation based on baseline survival probability and addressability, we performed an exploratory data analysis. This involved fitting the heuristic model to the optimal allocations for 120 tasks and comparing these with

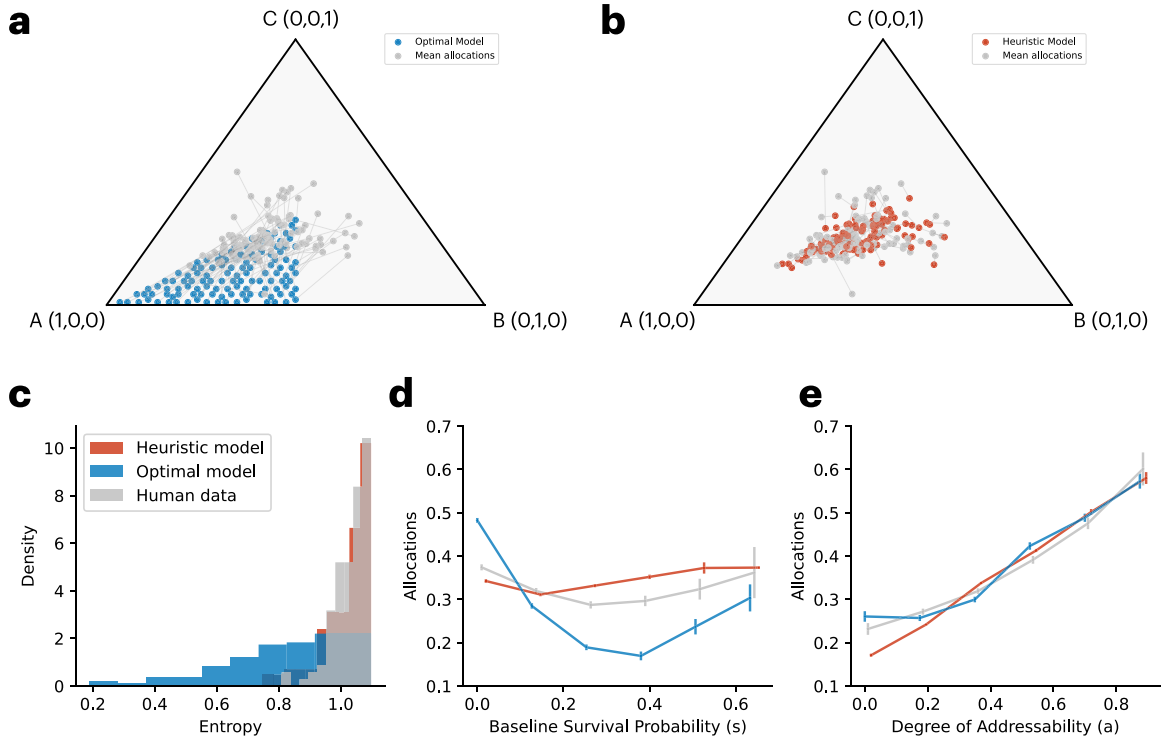

Fig. 2. Resource allocations to avoid existential risks in Experiment 1. (a) The optimal model (blue dots) compared with the mean normalized allocations made by participants (grey dots). Each vertex represents full investment in a given risk. Connecting lines indicate correspondence between model predictions and empirical data. (b) The heuristic model (red dots) compared with participants (grey dots). (c) Histogram of entropy values for allocations. (d) Mean allocation to a particular risk as a function of that risk's baseline survival probability. All else equal, risks with lower s_i should receive larger allocations, but this relationship is non-monotonic when averaging over our decision problems because those problems are generated using constraints that produce a correlation between s_i and a_i (see the supplement linked in Appendix A for details). (e) Mean allocation to a particular risk as a function of that risk's degree of addressability. Grey lines and bars represent empirical data, red ones represent the best-fitting heuristic model, and blue ones represent the optimal model predictions. Error bars denote 95% CI across individual allocations.

Table 1
Results of Experiment 1, comparing model predictions and empirical allocations in each of the bins for baseline survival probabilities (ordered from left to right).

Models	Test	1st	2nd	3rd	4th	5th	6th
Optimal model	t-test	−30.44**	9.12**	21.68**	18.23**	6.29**	1.80
	Cohen's d	−0.45	0.16	0.50	0.60	0.45	0.27
Heuristic model	t-test	0.52	4.61**	−3.27**	−3.54**	−1.79	0.38
	Cohen's d	0.01	0.08	−0.08	−0.12	−0.13	0.06

Note.
** Denotes statistically significant difference with $p < .01$.

Table 2
Results of Experiment 1, comparing model predictions and empirical allocations in each of the bins for addressability (ordered from left to right).

Models	Test	1st	2nd	3rd	4th	5th	6th
Optimal model	t-test	−3.15*	3.43**	5.28**	−6.35**	−1.81	1.38
	Cohen's d	−0.12	0.07	0.09	−0.13	−0.06	0.13
Heuristic model	t-test	−0.09	1.00	−2.21*	0.23	1.52	1.07
	Cohen's d	−0.00	0.02	−0.04	0.01	0.05	0.11

Note.
* Denotes statistically significant difference with $p < .05$
** Denotes statistically significant difference with $p < .01$.

empirical allocations. The analysis revealed that optimal allocations necessitate $\beta_s^{opt} = 3.63$ and $\beta_a^{opt} = 2.98$, which exceed the values fitted to human allocations by 2.66 and 0.59 respectively. Moreover, the heuristic model fitted to optimal allocations achieves a Pearson's $r = 0.87$ ($p < .01$), closely matching the correlation obtained when the model was fitted directly to people's allocations. This indicates that the heuristic can capture optimal allocations as well as it captures's people's allocations, with appropriate values for its parameters. The parameter values obtained from fitting people's judgments suggest that they generally do not respond sufficiently to either baseline survival probabilities or addressabilities, but respond less well to survival probabilities.

5. Experiment 2: Instruction variants

As mentioned above, the complacency with which a risk should be treated depends on the ratio of its baseline survival probability

to its addressability. Dependence on this ratio is a distinctive feature of multiplicative risk mitigation, and it is not obvious how the presentation of this ratio impacts decision-making in the particular case of existential risks. To investigate whether participants departed from optimal behavior simply because it was difficult to compute the survival-probability-to-addressability ratio (and in the light of known framing effects for different presentations of ratios (Larrick & Soll, 2008) as well as cases where explicit information about intervention effectiveness is frequently not taken to be decisive (Berman et al., 2018)), we conducted a second experiment in which people solved the same allocation problems with different framing conditions. In one condition, the problem statements emphasized survival probabilities. In the other condition, we calculated the ratio of survival probability to addressability and explicitly provided that ratio to participants.

5.1. Methods

5.1.1. Participants

We recruited 1201 and 1203 participants from Prolific respectively for the two distinct conditions in Experiment 2: (1) the survival probabilities condition and (2) the survival probabilities and multiplying factors condition. None of these participants had taken part in the first experiment, and they were restricted to participating in only one of the conditions. Participants were also required to be from the US and have a track record of at least 100 submissions with an acceptance rate of 95% or higher. Out of these, 797 participants (302 males, 307 females, 188 who chose not to disclose their gender, ages ranged between 18 and 88 with a median of 38) in the first condition and 492 (210 males, 175 females, 107 who chose not to disclose their gender, ages ranged between 18 and 80 with a median of 37) in the second successfully completed the experiment and passed the catch trial. Each participant was compensated with the same $3.00 base payment as was offered in Experiment 1. The experimental sessions were conducted in January 2024. All participants provided their consent in accordance with the Princeton University IRB number 10859 "Computational Cognitive Science".

5.1.2. Stimuli

We used the same set of 120 tasks as in Experiment 1.

5.1.3. Procedure

The tasks were presented using two distinct framings (preregistered at https://osf.io/536qs). In Experiment 1, the effect of investing in each disaster was described in terms of how much that investment would **reduce** the probability of that disaster **occurring**. In contrast, in each of the two conditions of Experiment 2 the effect of investing in each disaster was instead described in terms of how much that investment would **increase** the probability of that disaster **being avoided**.

The two conditions differed in how we conveyed the magnitude of the effect of investment in a given disaster. In the first condition (illustrated in Figure A6), we characterized the profile of disaster i by way of three pieces of information: the baseline probability of avoiding the disaster ($s_i(0)$), the addressability (a_i), and the probability of avoiding the disaster if the entire budget were to be allocated towards its mitigation ($s_i(T)$). In the second condition (illustrated in Figure A7), for each disaster we also prominently included the factor that the probability of avoiding that disaster would be multiplied by if the entire budget were allocated to it ($s_i(T)/s_i(0)$).

5.2. Results

The results of this experiment are shown in Figs. 3 and 4. We found that the observed systematic departures from optimal behavior persisted. Empirical allocations in these two conditions were highly correlated with our original results ($r = 0.83, p < .01$; $r = 0.90, p < .01$). Mean allocations also showed no statistically significant differences, with t-tests yielding $t(359) = 0.00 (p = 1.00)$.

Employing a multiplying factor in the second condition resulted in a higher failure rate in the catch trial (59.10%) compared to 33.64% in the first condition and 34.70% in Experiment 1. This indicates that participants had greater difficulty interpreting the multiplying factors when making their allocations. Despite restricting the sample to participants who passed the catch trial, indicating a nominal understanding of the task, the allocations produced by those participants still significantly diverged from the optimal model and showed high correlation across all three framing conditions.

We applied the heuristic model to the data from each of the two conditions independently. For the first condition, the best-fitting parameters were $\beta_s = 0.34, \beta_a = 2.94$, while for the second condition, they were $\beta_s = 0.40, \beta_a = 3.31$. In terms of correlation between data and model predictions, the best-fitting heuristic model achieved a Pearson's $r = 0.92$ ($p < .01$) for the first condition and $r = 0.87$ ($p < .01$) for the second. By comparison, the optimal model only managed to achieve a Pearson's $r = 0.52$ ($p < .01$) for both conditions. The adjusted R^2 values of the heuristic models in capturing empirical allocations are 0.84 and 0.75 for the first and second conditions, respectively, whereas the optimal model yields substantially lower adjusted R^2 values of 0.27 for both conditions. Figs. 3 and 4 illustrate these findings.

We also conducted statistical analyses to determine the impact of framing on observed resource allocation in mitigating existential risks. The findings, summarized in Table 3, are based on Pearson's r and paired t-tests. To delve deeper into the effects of framing, we executed a two-way ANOVA, taking empirical allocations as the dependent variable and incorporating framing, an indicator of the best-fitting model, and their interaction as independent variables. Specifically, empirical allocations were analyzed at the task level. For each of the 120 unique tasks, average allocations were computed for each of the three existential risks. Therefore, for each framing condition, a total of $120 \times 3 = 360$ data points served as the dependent variable in our analyses. The ANOVA results revealed that the effect of framing $F(2358) = 0.0031$ ($p = 1.00$), the best-fitting model $F(1359) = 0.5248$ ($p = 0.47$), and the interaction between framing and the best-fitting model $F(2358) = 1.1762$ ($p = 0.31$) were all not statistically significant. These results imply a high correlation in empirical allocations. Therefore, we cannot reject the null hypothesis, which posits no difference in empirical allocations across various framings.

Because the three allocations are constrained to sum to 1, we conducted a Dirichlet regression to account for this dependency structure. Mirroring the prior ANOVA, we used the mean task-level allocations as the dependent variable, and included framing condition, best-fitting model, and their interaction as predictors. The regression coefficients for framing condition, best-fitting model, and their interaction were 0.0420 ($p = 0.65$), −0.1355 ($p = 0.70$), and 0.1449 ($p = 0.72$), respectively. These results do not indicate any statistically significant effects on the empirical evaluations from framing condition, best-fitting model, or their interaction.

The same statistical tests for marginal effects of baseline survival probability and addressability from the previous experiment were applied to the two new conditions. Results of the t-tests and Cohen's d are presented in Table 4 and 5 for the first condition, and in Table 6 and 7 for the second. Similar to the first experiment, the optimal model had a poor fit, diverging in 20 of 24 bins across both conditions at a statistically significant level for $p < .05$, while the best-fitting heuristic model diverged in only 9 of 24 bins.

6. Discussion

Our results reveal the optimal strategy for solving the distinctive multiplicative decision problems that are involved in mitigating existential risks and suggest that people may be misled by their intuitions when thinking about these problems. These results complement both

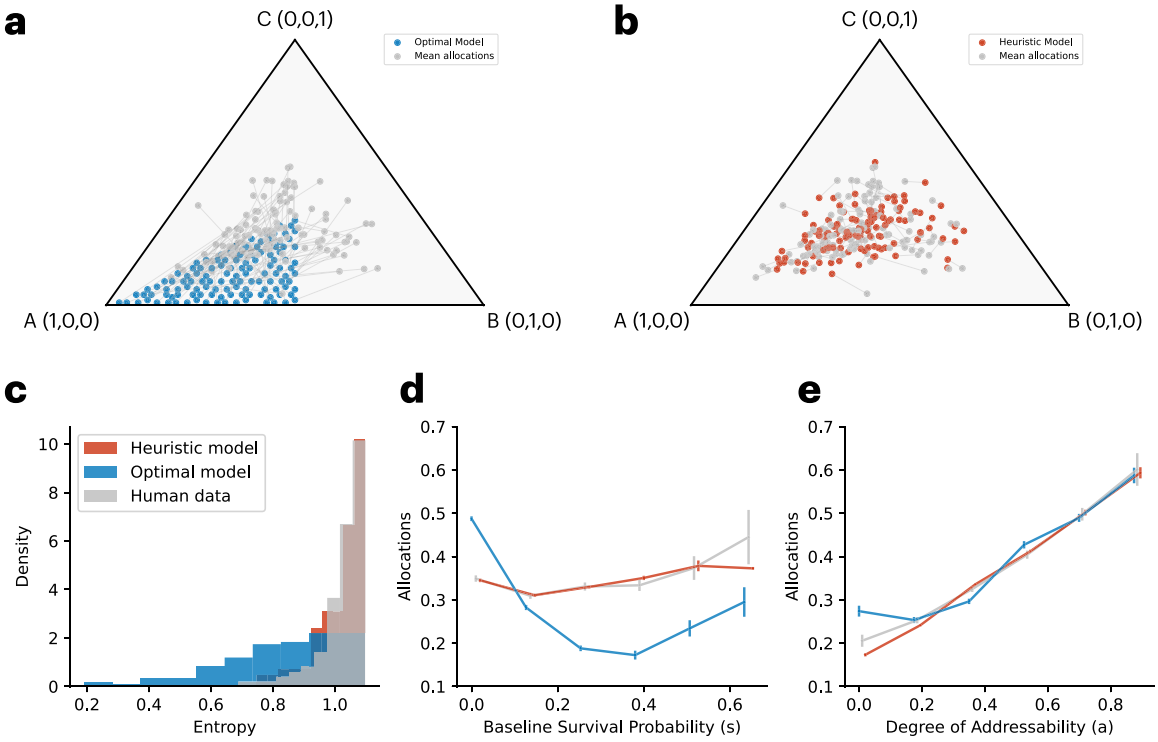

Fig. 3. Results of Experiment 2 (survival probability framing). Comparison of model predictions with the mean normalized allocations made by participants (grey dots). Each vertex represents full investment in a given risk. Connecting lines indicate correspondence between model predictions and empirical data. (a) The optimal model (blue dots). (b) The heuristic model (red dots). (c) Histogram of entropy values for allocations. (d) Mean allocation to a particular risk as a function of that risk's baseline survival probability. (e) Mean allocation to a particular risk as a function of that risk's degree of addressability. Grey lines and bars represent empirical data, red ones represent the best-fitting heuristic model, and blue ones represent the optimal model predictions. Error bars denote 95% CI across individual allocations.

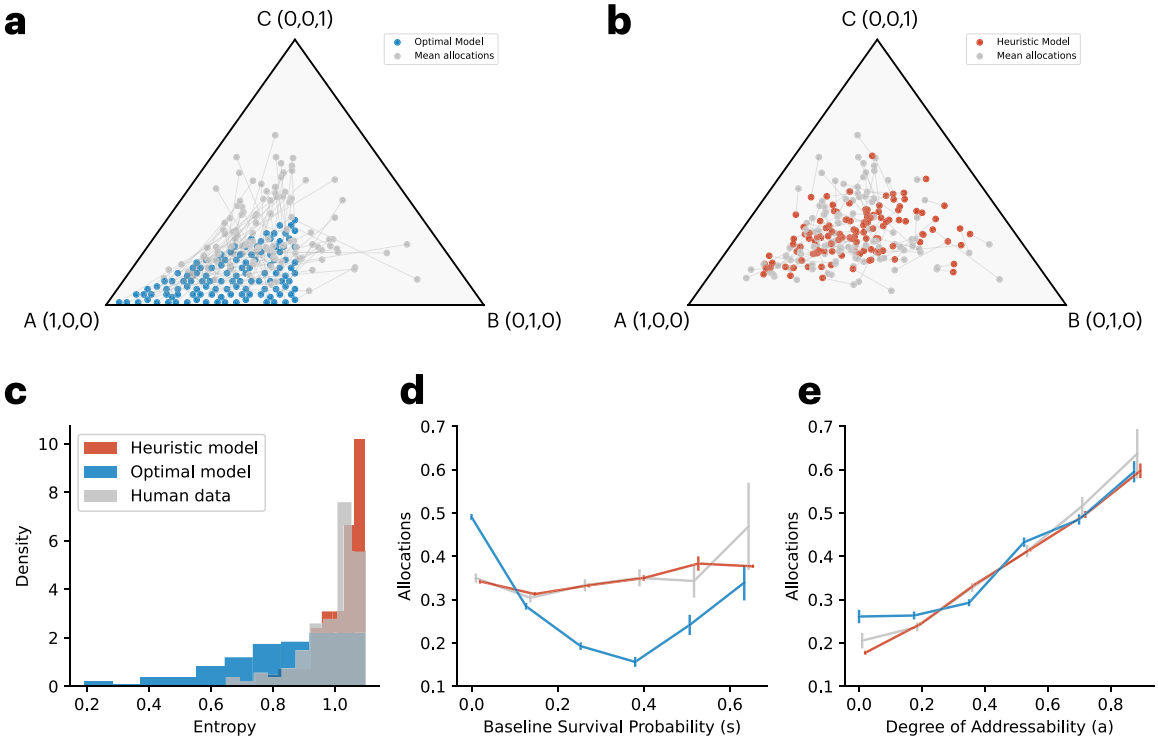

Fig. 4. Results of Experiment 2 (the survival probabilities and multiplying factors condition). Comparison of model predictions with the mean normalized allocations made by participants (grey dots). Each vertex represents full investment in a given risk. Connecting lines indicate correspondence between model predictions and empirical data. (a) The optimal model (blue dots). (b) The heuristic model (red dots). (c) Histogram of entropy values for allocations. (d) Mean allocation to a particular risk as a function of that risk's baseline survival probability. (e) Mean allocation to a particular risk as a function of that risk's degree of addressability. Grey lines and bars represent empirical data, red ones represent the best-fitting heuristic model, and blue ones represent the optimal model predictions. Error bars denote 95% CI across individual allocations.

Table 3
Comparing empirical allocations between Experiment 1 and the two conditions of Experiment 2.

Test	Exp 1 vs. the 1st cond.	Exp 1 vs. the 2nd cond.	The 1st vs. the 2nd cond.
Pearson's r	$r = 0.83$ ($p < .01$)	$r = 0.90$ ($p < .01$)	$r = 0.85$ ($p < .01$)
Paired t-test	$t(359) = 0.00$ ($p = 1.00$)	$t(359) = 0.00$ ($p = 1.00$)	$t(359) = 0.00$ ($p = 1.00$)

Table 4
Results for Experiment 2 (survival probabilities condition), comparing model predictions and empirical allocations in each of the bins for baseline survival probabilities (ordered from left to right).

Models	Test	1st	2nd	3rd	4th	5th	6th
Optimal model	t-test	−35.56**	6.94**	27.76**	21.17**	8.50**	4.05**
	Cohen's d	−0.56	0.12	0.68	0.72	0.64	0.69
Heuristic model	t-test	2.09*	−1.24	−0.18	−2.64**	−0.10	2.40*
	Cohen's d	0.03	−0.021	−0.00	−0.09	−0.01	0.40

Note.
* Denotes statistically significant difference with $p < .05$.
** Denotes statistically significant difference with $p < .01$.

Table 5
Results for Experiment 2 (survival probabilities condition), comparing model predictions and empirical allocations in each of the bins for addressability (ordered from left to right).

Models	Test	1st	2nd	3rd	4th	5th	6th
Optimal model	t-test	−7.30**	−0.05	6.77**	−3.98**	1.09	0.71
	Cohen's d	−0.27	−0.00	0.12	−0.09	0.04	0.07
Heuristic model	t-test	2.71**	0.53	−3.77**	0.09	2.06*	1.99*
	Cohen's d	0.10	0.01	−0.07	0.00	0.07	0.19

Note.
* Denotes statistically significant difference with $p < .05$.
** Denotes statistically significant difference with $p < .01$.

Table 6
Results for Experiment 2 (survival probabilities and multiplying factors condition), comparing model predictions and empirical allocations in each of the bins for baseline survival probabilities (ordered from left to right).

Models	Test	1st	2nd	3rd	4th	5th	6th
Optimal model	t-test	−25.86**	3.59**	18.85**	17.98**	4.43**	2.49*
	Cohen's d	−0.53	0.08	0.58	0.80	0.46	0.50
Heuristic model	t-test	1.62	−1.81	0.02	0.04	−2.30*	1.83
	Cohen's d	0.03	−0.04	0.00	0.00	−0.19	0.38

Note.
* Denotes statistically significant difference with $p < .05$.
** Denotes statistically significant difference with $p < .01$.

Table 7
Results for Experiment 2 (survival probabilities and multiplying factors condition), comparing model predictions and empirical allocations in each of the bins for addressability (ordered from left to right).

Models	Test	1st	2nd	3rd	4th	5th	6th
Optimal model	t-test	−4.75**	−4.34**	5.78**	−2.43*	2.84**	1.35
	Cohen's d	−0.22	−0.12	0.14	−0.07	0.12	0.18
Heuristic model	t-test	3.19**	−1.37	−1.62	−0.54	1.95	1.32
	Cohen's d	0.14	−0.04	−0.04	−0.02	0.08	0.18

Note.
* Denotes statistically significant difference with $p < .05$.
** Denotes statistically significant difference with $p < .01$.

previous work exploring multiplicative decisions with more graded forms of risk (Brennan & Xia, 2002; Campbell & Viceira, 2001; Franke et al., 2011) and longstanding findings of suboptimal decision-making in additive decision problems (Tversky & Kahneman, 1974).

Our mathematical analysis shows that decisions about mitigating existential risks can be reduced to evaluating a simple quantity — the ratio of survival probability to addressability. Our behavioral experiments, however, indicate that people struggle to optimally use this quantity when making investment decisions to mitigate existential risks. Even when the ratio of survival probability to addressability is explicitly presented in the stimuli (i.e., the multiplying factors condition of Experiment 2), participants did not produce improved allocations compared to situations where that ratio had to be calculated. In other words, the three tested frames had minimal impact on nudging participants towards better solutions. These findings suggest that people exhibit a strong tendency to rely on heuristics to make decisions about existential risks, heuristics that are better suited to decision problems with additive rather than multiplicative incentives. Knowing about this tendency matters for two main reasons. First, it points to the unexpectedly high value of educating decision-makers on how to compare existential risk interventions. Second, our results suggest that an extremely natural and seemingly obvious practice – presenting risk mitigation options individually for comparison – is suboptimal. Put bluntly: while people are relatively good at choosing individual items for their grocery carts given each item's price per ounce (a linear allocation problem), people are less good at the nonlinear allocation problem of choosing individual existential risk interventions based on

each intervention's risk reduction per dollar spent. As a result, it may be better to instead present decision makers with competing *total allocation profiles* (specifying how much is to be invested in each risk), where each profile is explicitly labeled with its estimated reduction in *overall* existential risk (the risk that any catastrophe occurs). Since choosing among such profiles would not require decision makers to directly compare individual interventions, framing the allocation problem this way would mitigate the allocation distortions we have described above. Corresponding reframings could improve choices in other decision problems that involve multiplicative incentives.

There are a number of ways in which this work could be extended to explore the impact of other factors that have been shown to be relevant to human decision-making. For example, providing a concrete context in which the scenarios are presented (c.f. Cosmides & Tooby, 1992) and clarifying the underlying causal structure of events (c.f. Krynski & Tenenbaum, 2007), rather than just presenting abstract risks characterized by addressability and survival probability may influence the strategy that people use. Another relevant consideration is temporal discounting (Doyle, 2013): if risks of extinction operate on different timescales, people may be willing to focus more on more imminent risks regardless of their probability. Exploring these questions is essential if we want to develop a more complete understanding of how people make decisions about existential risks.

We acknowledge several limitations of our current approach, which warrant further investigation. First, while we explored laypeople's intuitions about decisions involving existential risks, these results should be replicated with policymakers if we want to make stronger recommendations for how societies should engage with these decisions. Second, while assuming a linear return on investment to reduce risk probabilities simplified the calculation of the optimal solution and was explicitly presented in our experiment, this assumption is unrealistic in most real-world scenarios. In practice, people often encounter diminishing returns, where the benefit from additional investment decreases as the investment amount increases. This can be accommodated through a modification to our mathematical approach, at the cost of some additional complexity. Third, although the framing effect was found to be non-significant in our experiments, there are many alternative ways to convey information about mitigation projects. Future research should explore how to better nudge decision-making when the risk structure is multiplicative rather than additive. Lastly, the underlying reasons why participants consistently adopted the specific heuristics we identified, across different framing conditions, remain unclear. Theoretical work using resource-rational analysis (Lieder & Griffiths, 2020) may shed light on why these particular heuristics prevail when dealing with multiplicative risks.

CRediT authorship contribution statement

Adam Elga: Writing – review & editing, Writing – original draft, Supervision, Software, Formal analysis, Conceptualization. **Jian-Qiao Zhu:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Thomas L. Griffiths:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Acknowledgments

This work and related results were made possible with the support of the NOMIS Foundation. We thank Ben Enke, Sanjeev Kulkarni, Danny Oppenheimer, and Pietro Ortoleva for helpful discussions. A. Elga thanks Mindy and Gene Stein for research environment support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106216.

Data availability

Links to data are included in paper.

References

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, *9*(3), 396–406. http://dx.doi.org/10.1016/0030-5073(73)90061-5.

Benartzi, S., & Thaler, R. H. (2001). Naive diversification strategies in defined contribution saving plans. *American Economic Review*, *91*(1), 79–98.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., et al. (2023). Managing AI risks in an era of rapid progress. arXiv preprint arXiv:2310.17688.

Berman, J. Z., Barasch, A., Levine, E. E., & Small, D. A. (2018). Impediments to Effective Altruism: The Role of Subjective Preferences in Charitable Giving. *Psychological Science*, *29*(5), 834–844, arXiv:26957434.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31.

Brennan, M. J., & Xia, Y. (2002). Dynamic asset allocation under inflation. *The Journal of Finance*, *57*(3), 1201–1238.

Campbell, J. Y., & Viceira, L. M. (2001). Who should buy long-term bonds? *American Economic Review*, *91*(1), 99–127.

Carlsmith, J. (2023). Existential risk from power-seeking AI. In J. Barrett, H. Greaves, & D. Thorstad (Eds.), *Essays on longtermism*. Oxford: Oxford University Press.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 163–228). Oxford University Press.

Cotton-Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: prevention, response, resilience, and why they all matter. *Global Policy*, http://dx.doi.org/10.1111/1758-5899.12786, 10.1111/1758–5899.12786.

Critch, A., & Russell, S. (2023). TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. http://dx.doi.org/10.48550/arXiv.2306.06924, arXiv:2306.06924.

Doyle, J. R. (2013). Survey of time preference, delay discounting models. *Judgment and Decision Making*, *8*(2), 116–135.

Dutt, V. (2013). Responding Linearly in Nonlinear Problems: Application to Earth's Climate. In M. Carpenter, & E. J. Shelton (Eds.), *Environmental remediation technologies, regulations and safety*, *Carbon dioxide emissions: new research*. New York: Nova Publishers.

Fan, Y., Budescu, D. V., & Diecidue, E. (2019). Decisions with compound lotteries. *Decision*, *6*(2), 109–133. http://dx.doi.org/10.1037/dec0000091.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*(3), 195–200.

Franke, G., Schlesinger, H., & Stapleton, R. C. (2011). Risk taking with additive and multiplicative background risks. *Journal of Economic Theory*, *146*(4), 1547–1568.

Greaves, H., & MacAskill, W. (2021). *The case for strong longtermism: GPI working paper no. 5–2021*, Global Priorities Institute.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430–450.

Larrick, R. P., & Soll, J. B. (2008). The MPG Illusion. *Science*, *320*(5883), 1593–1594.

Lewis, J., Feiler, D., & Adner, R. (2023). The Worst-First Heuristic: How Decision Makers Manage Conjunctive Risk. *Management Science*, *69*(3), 1575–1596. http://dx.doi.org/10.1287/mnsc.2022.4411.

Lewis, J., & Simmons, J. P. (2020). Prospective outcome bias: Incurring (unnecessary) costs to achieve outcomes that are already likely. *Journal of Experimental Psychology: General*, *149*(5), 870–888. http://dx.doi.org/10.1037/xge0000686.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, Article e1.

MacAskill, W. (2023). *What we owe the future*. New York: Basic Books.

Mckenzie, C. R., & Liersch, M. J. (2011). Misunderstanding Savings Growth: Implications for Retirement Savings Behavior. *Journal of Marketing Research (JMR)*, *48*, S1–S13. http://dx.doi.org/10.1509/jmkr.48.SPL.S1.

Millett, P., & Snyder-Beattie, A. (2017). Existential risk and cost-effective biosecurity. *Health Security*, *15*(4), 373–383. http://dx.doi.org/10.1089/hs.2017.0028.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Nilsson, H., Rieskamp, J., & Jenny, M. (2013). Exploring the Overestimation of Conjunctive Probabilities. *Frontiers in Psychology*, *4*, http://dx.doi.org/10.3389/fpsyg.2013.00101.

Olsson, A.-C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(6), 1371–1384. http://dx.doi.org/10.1037/0278-7393.32.6.1371.

Ord, T. (2020). *The precipice: existential risk and the future of humanity*. New York: Hachette Books.

Savage, L. J. (1954). *Foundations of statistics*. New York: John Wiley & Sons.

Schubert, S., Caviola, L., & Faber, N. S. (2019). The Psychology of Existential Risk: Moral Judgments about Human Extinction. *Scientific Reports*, *9*(1), 1–8. http://dx.doi.org/10.1038/s41598-019-50145-9.

Slovic, P. (2020). Risk perception and risk analysis in a hyperpartisan and virtuously violent world. *Risk Analysis*, *40*(S1), 2231–2239.

Thorstad, D. (2023). High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation. *Philosophy and Public Affairs*, *51*(4), 373–412.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131.

Van Dooren, W., De Bock, D., Depaepe, F., Janssens, D., & Verschaffel, L. (2003). The Illusion of Linearity: Expanding the evidence towards probabilistic reasoning. *Educational Studies in Mathematics*, *53*(2), 113–138. http://dx.doi.org/10.1023/A:1025516816886.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Voosen, P. (2023). Earth may face higher risk of catastrophic asteroid strikes. *Science*, *379*(6638), http://dx.doi.org/10.1126/science.adh9058, 1179–1179.

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, *107*(1), 105–136. http://dx.doi.org/10.1016/j.cognition.2007.08.003.