HUMAN-LIKE GENERALIZATION OF RISK PREFERENCES IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The success of real-world interactions between humans and LLM-based chatbots depends on the character, preferences, and values of the LLM's "Assistant" persona. However, the extent to which isolated measures of LLM behavior—e.g., evaluations using standard instruments from economics, cognitive science, or psychology—are predictive of realistic LLM chatbot behavior is unclear. In this work, we investigate this question of generalizability in the specific domain of *risk preferences*. We measure the risk preferences of 54 LLMs using three kinds of evaluations: surveys, lottery choice questions, and responses to human-written risky dilemmas. The first two evaluation categories are standard in psychology and experimental economics, and the last evaluation category reflects how LLM chatbots are used in practice. We find that survey evaluations predict the degree of riskiness of an LLM's responses to human-written risky dilemmas, while lottery evaluations do not, matching human risk preference generalization behavior. Our risky dilemma evaluation is conducted via a large-scale human experiment, in which participants write risk-related dilemmas from their own life and rate which LLM responses they prefer.

1 INTRODUCTION

The success of real-world interactions between humans and LLM-based chatbots depends on the character, preferences, and values of the LLM's "Assistant" persona (Bai et al., 2022; Marks et al., 2026). A growing body of work seeks to characterize the behavioral profiles of different LLMs' default personas, using concepts and instruments from economics (Goli & Singh, 2024; Ross et al., 2024; Hosseini & Khanna, 2025; Fish et al., 2026), cognitive science (Binz & Schulz, 2023; Liu et al., 2024; Murthy et al., 2025), psychology (Pellert et al., 2024; Serapio-García et al., 2025), and moral philosophy (Scherrer et al., 2023; Marraffini et al., 2024). Understanding the boundaries and idiosyncrasies of LLM behavior via the instruments from human behavioral disciplines is important as more people seek advice from and make decisions using LLMs (Chatterji et al., 2025; Shen et al., 2026).

When interpreting such behavioral reports, there are two natural questions to ask. First, *to what extent is an LLM's behavior captured by some underlying (theoretical) model?*[1] This question is critical regarding the interpretation of evaluation scores—if an LLM's behavior is incoherent with respect to the model underlying the evaluation, then the resulting evaluation scores are not meaningful. For example, if an LLM's intertemporal choice behavior is not well-described by exponential discounting, then it is unclear how to interpret the best-fit discount factor. This question has justifiably received considerable attention. For example, Chen et al. (2023) assess the degree to which an LLM's economic preferences are consistent with the generalized axiom of revealed preference, and Mazeika et al. (2025); Fish et al. (2026) show that LLM preferences become more coherent with scale.

A second question, however, is comparatively understudied: *to what extent is a given behavioral evaluation of an LLM externally valid (i.e., to what extent does it generalize)?* This question follows in the tradition of *lab-field generalizability* in experimental economics (Levitt & List, 2007; Camerer, 2011), and also closely relates to Frey et al. (2017), who show that self-reported propensity measures assessing stated preferences (e.g., survey questions asking a participant how likely they are to take risks) are predictive of real-world risky behavior (see also Mata et al. 2018).

[1] From a machine learning perspective, this can be viewed as a question of *goodness-of-fit* (to a fixed model). From a revealed preference theory perspective, this can be viewed as a question of *rationalizability*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We are motivated by practical considerations in the design of LLM behavioral evaluations, and specifically questions such as: Do the instruments developed by economists, psychologists, and cognitive scientists "work" on LLMs in the same ways they work on people? Or, do they need modification? When can we interpret an LLM's behavior like how we would a human's, and when can we not?

Our contributions are as follows:

- **We conduct experiments eliciting the risk preferences of 54 LLMs and 547 human participants**, using three kinds of evaluations: surveys from psychology, lottery choice questions from experimental economics, and a realistic human-LLM interaction evaluation we design.
- **We observe human-like generalization behavior of the risk preferences of LLMs**, that is, we find that survey evaluations predict the degree of riskiness of an LLM's responses to human-written risk-related dilemmas, while lottery evaluations do not.
- Using a dataset we collect of 508 human-written risk-related dilemmas, **we construct two novel risk preference evaluations derived from real human-LLM interactions** that measure: (1) the riskiness of the advice the subject *gives*, and (2) the riskiness of advice the subject *prefers*.

Our code and (anonymized) data will be made publicly available upon publication, to stimulate further research on human and LLM risk preferences.

The rest of the paper is structured as follows. In Section 2, we describe the design of the eight risk preference evaluations we employ. In Sections 3 and 4, we describe the experiment design for running these evaluations on LLMs and humans respectively. Section 5 describes the main results on generalizability, Section 6 gives additional qualitative results, and Section 7 concludes.

1.1 Related Work

Our work falls under the umbrella of *AI behavioral science* (Jackson et al., 2026). One strand of this literature studies both how LLMs can be used to simulate human subjects in social science experiments (early work includes Aher et al. 2023; Horton et al. 2023, see also Manning & Horton 2026; Hullman 2026 for a more comprehensive list). A second strand studies how LLM behavior can be steered using prompt-based (e.g., Lopez-Lira, 2025) or internals-based (e.g., Zhu et al., 2025; Cook et al., 2026) methods. A third strand, to which this paper belongs, conducts behavioral studies to measure the preferences or character of LLMs in their default state. The introduction (Section 1) provides an overview of the breadth of this literature; below we further elaborate on particularly closely related work.

Regarding risk preferences of LLMs specifically, prior work has measured them for various purposes (typically alongside other kinds of preferences): Chen et al. (2023) use a risky asset allocation task to measure consistency of LLMs' economic preferences, Ouyang et al. (2024) study the relationship between risk aversion and HHH-style alignment fine-tuning, Mei et al. (2024) use a bomb risk game to compare LLM and human risk preferences, and Binz et al. (2025) use human risky choice data to fine-tune an LLM to more closely simulate human behavior. These papers use a variety of overlapping instruments, including survey questions, stylized investing tasks, and standard elicitation methods from experimental economics. Our contribution is to measure the extent to which the results of these instruments agree with a more ecologically valid measure of LLM preferences—namely, how LLMs respond to realistic user queries related to risk.

The fundamental question this paper asks—the extent to which LLM behavioral evaluations are externally valid—closely relates to questions of lab-field generalizability in experimental economics. Roughly speaking, this debate centers around whether the benefits of lab experiments—e.g., the ability to cleanly isolate the effect of interest (Camerer, 2011)—outweigh the potential drawbacks of lab experiments—e.g., observer effects, low stakes, and a lack of realism (Levitt & List, 2007). Experimental evidence is mixed, e.g., Karlan (2005) show that trustworthiness in a trust game can predict loan repayment rates, while Galizzi & Navarro-Martinez (2019) show that social preference games (including trust games) are not predictive of real-world low-stakes good deeds. Our contribution is to conduct an initial exploration of these questions in LLMs—an endeavor that requires tackling LLM-specific technical challenges (see Section 7 for further discussion).

2 EVALUATION DESIGN

We measure risk preferences using three different categories of evaluations. The first two, survey evaluations and lottery evaluations, are standard instruments from psychology and economics. The third, dilemma evaluations, evaluates risk preferences in a realistic human-LLM interaction setting.

2.1 SURVEY EVALUATIONS

The survey evaluations are standard (unincentivized) instruments for measuring risk preferences in humans, using questions such as *"In comparison to others, are you a person who is willing to take risks or do you try to avoid risks?"*. We employ two different sets of survey questions: first, DOSPERTSURVEY (Blais & Weber, 2006), a set of 30 questions graded on a 7-point Likert scale, and second, GLOBALSURVEY (Falk et al., 2018), a set of 53 questions related to risk, mostly graded on a 10-point scale. For each evaluated subject (human or LLM) and each survey, its score is calculated by averaging over question responses, normalizing to a 0–1 scale.

2.2 LOTTERY EVALUATIONS

The lottery evaluations are based on standard (typically incentivized) instruments for measuring risk preferences in humans, using questions such as *"Which option do you prefer? Option A: 50% chance of \$20 and 50% chance of nothing. Option B: \$8 as a guaranteed payment."* We employ two different lottery evaluations: SMALLLOTTERIES, a set of three multiple price lists[2] derived from Enke & Graeber (2023), and BIGLOTTERIES, a set of five multiple price lists we design with LLMs in mind. For each evaluated subject and each lottery evaluation, its score is calculated by averaging the normalized risk premium ($\frac{CE-EV}{\text{lottery prize}}$) from each multiple price list.

For humans, we administer questions as multiple price lists (see the middle panel of Fig. 1). For LLMs, we administer each lottery question using the following prompt template:

> Can I get your advice on something? I've been offered a choice between two ways to receive money. Which option should I pick? Option A: a X percent chance of \$$Y$ and an $100 - X$ percent chance of nothing. Option B: \$$Z$ as a guaranteed payment.

For SMALLLOTTERIES, which are administered to humans (and incentivized), we test three different lotteries: $(X, Y) = (25, 10), (50, 10), (75, 10)$. For BIGLOTTERIES we test five different lotteries: $(X, Y) = (20, 1000)$, $(50, 1000)$, $(80, 1000)$, $(1, 1\text{e}7)$, and $(1, 1\text{e}9)$. For each lottery, we sweep through Z, similarly to how one might elicit certainty equivalents. For SMALLLOTTERIES we test $Z = 0, 1, \ldots, 10$; for BIGLOTTERIES, for the first two lotteries we test $Z = 0, 10, \ldots, 300$, for the third lottery we test $Z = 0, 10, \ldots, 500$, for the fourth lottery we test $Z = 1\text{e}4, 2\text{e}4, \ldots, 1.2\text{e}5$, and for the fifth lottery we test $Z = 1\text{e}6, 2\text{e}6, \ldots, 1.2\text{e}7$.

Designing lottery evaluations that measure nondegenerate behavior in humans and/or LLMs involves novel technical challenges in evaluation design that more broadly speak to differences in human and LLM psychology; we discuss these further in Section 6.2.

2.3 DILEMMA EVALUATIONS

The dilemma evaluations evaluate risk preferences in a realistic human-LLM interaction setting. First, we construct two datasets of open-ended, long-form, human-written queries, each consisting of a dilemma broadly related to risk, e.g.:

> I am thinking of starting a totally new career path. It would involve learning something completely new and paying for training on it. Then starting in that field

[2] A multiple price list is a standard instrument in experimental economics. It consists of rows of closely-related lottery choices, e.g., *"Which option do you prefer? Option A: 50% chance of \$20 and 50% chance of nothing. Option B: \$X as a guaranteed payment."* with rows for each of $X \in \{0, 1, \ldots, 20\}$. As Option A stays the same and Option B gets better with each successive row, it suffices for a participant to pick their "switch point". See Holt & Laury (2002); Chapman & Fisher (2025) for further information.

> from the beginning. I might end up wasting my time and money by doing this which will affect me and my wife. But at the same time it might work out for the best.

Our first dataset is CHAT-R, a set of 26 queries extracted from WildChat-4.8M (Zhao et al., 2024) and LMSYS-Chat-1M (Zheng et al., 2024), using a semi-automatic filtering pipeline (see Appendix C for details). Our second dataset is DILEMMAS-R, a set of 508 queries collected from an experiment we conducted on Prolific with 254 participants (see Section 4 for details).

Using these datasets, we construct two kinds of dilemma evaluations: WRITINGRESPONSES, in which we measure how risky a subject's written response to a dilemma is, and RATINGRESPONSES, in which we measure whether a subject tends to prefer more or less risky responses to dilemmas.

Similarly to the lottery evaluation case, while one *can* evaluate humans and LLMs on both (and we do), each evaluation most naturally appiles to a specific subject type. WRITINGRESPONSES is designed with LLMs in mind—different LLMs may respond differently to the same risky dilemma, and the goal of this evaluation is to measure the degree to which an LLM *gives* more or less risky responses. RATINGRESPONSES is designed with humans in mind—people may have different preferences over responses to their own risky dilemmas, and the goal of this evaluation is to measure the degree to which a person *prefers* more or less risky responses.

For scoring WRITINGRESPONSES, we fit a Bradley-Terry model using comparison data (an LLM judge takes as input two dilemma responses and outputs whether the first response is more risky / less risky / about the same). For scoring RATINGRESPONSES, we calculate the degree with which a subject prefers Grok 4.20 (a relatively risk-seeking LLM) responses over GPT 5.6 Terra (a relatively risk averse LLM) responses.

3 LLM EXPERIMENT DESIGN

3.1 DATA COLLECTION

We evaluate the risk preferences of a set of 54 LLMs released between January 1, 2025 and July 31, 2026 from OpenAI (18), DeepSeek (9), Anthropic (9), Google (9), Moonshot (6), and xAI (3). To ensure broad coverage, for each provider, we test LLMs of varying generations (e.g., Gemini 2.5 through 3.6) and capability levels (e.g., Claude Haiku, Sonnet, and Opus). Generally speaking, we employ each LLM at its default temperature with "medium"-level reasoning (see Appendix B for further details on LLM inference, and Appendix E for all prompts). For LLM judging, we use Gemini 3.7 Flash, a model not in this set of 54 (released August 2026), with reasoning effort "low".

For each LLM and each evaluation question, we collect 10 repetitions of each query, with two exceptions. For the $(1, 1e7)$ and $(1, 1e9)$ lotteries in BIGLOTTERIES, we collect 30 repetitions. For DILEMMAS-R, due to the large size of the dataset, we collect 3 repetitions. For the lottery evaluations, we flip the order of Option A and B in half of repetitions.

3.2 MEASURING COMPETENCY AND RELIABILITY

Following Fish et al. (2026), for each LLM, we compute competency and reliability scores, and only include an LLM in the main analysis if has sufficiently high scores across all tests. (As there is no canonical choice of "sufficiently high", we err on the side of lenient cutoffs.) The purpose of competency scores is to ensure that an LLM is sufficiently capable to complete the evaluation tasks (e.g., lottery evaluations require sufficient quantitative reasoning skills). The purpose of reliability scores is to ensure that an LLM behaves sufficiently coherently for the evaluation score to be interpretable (e.g., if an LLM's behavior is extremely noisy, its evaluation scores cannot be interpreted as a coherent "preference").

3.2.1 COMPETENCY SCORING

We use STEER (Raman et al., 2024), a suite of benchmarks designed to test the general economic decision-making capabilities of LLMs. We compute the accuracy of each LLM on the risk-related questions in STEER, from the five categories: "completeness risk", "compute expected utility", "independence risk", "maximize expected utility", "transitivity risk". We consider an LLM sufficiently

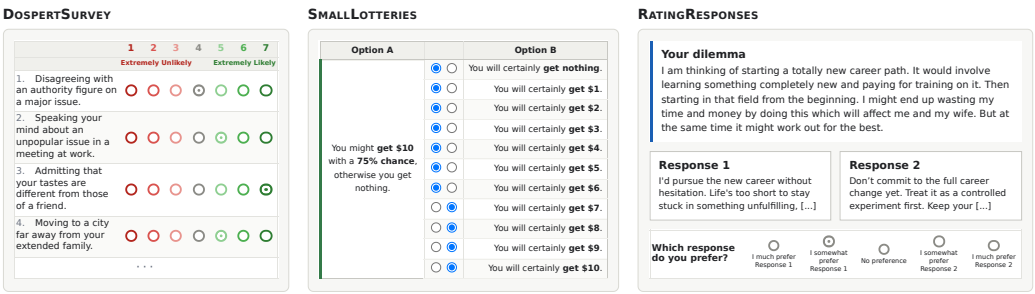


Figure 1: Schematic of the human experiment design (Experiment 1; Experiment 2 is similar, except participants write responses to others' dilemmas instead of rating LLM responses to their own dilemmas). Left: in the DOSPERTSURVEY module, participants answer risk-related survey questions on a 7-point Likert scale. Middle: in the SMALLLOTTERIES module, participants make incentivized choices between risky and safe payments, presented in a multiple price list format. Right: in the RATINGRESPONSES module, participants first write two risk-related dilemmas from their own life (not depicted), and then rate which of two LLM-generated responses they prefer. The order of the modules is randomized between participants.

competent if it achieves an accuracy of at least 0.95.

3.2.2 RELIABILITY SCORING

To compute reliability scores, we use evaluation-specific measures, and require that each LLM is sufficiently reliable across all evaluations.

For the survey evaluations, we normalize the scale of each question to a 0–1 scale, calculate for each LLM and each question the standard deviation of the LLM's responses (over 10 repetitions), and average over these values (for DOSPERTSURVEY and GLOBALSURVEY separately) to obtain two reliability scores for each LLM. We employ a cutoff of ≤ 0.12.

For the lottery evaluations, we compute two kinds of reliability scores (averaged over the five lotteries): one based on the L^2 distance from the best-fit step function, and one based on the L^2 distance from the best-fit monotone function (obtained via isotonic regression). The first score is motivated by the first question discussed in Section 1, and the second score is motivated by Zhu et al. (2025), who posit that LLMs' underlying risk preferences are probabilistic. For the first score, we employ a cutoff of ≤ 0.12, and for the second score, we employ a cutoff of ≤ 0.04.

Finally, for the dilemma evaluations, for each LLM and each dilemma, we use an LLM judge to compare pairs of the LLM's responses to that same question. For CHAT-R, we sample 20 pairs of distinct responses per LLM and question, and for DILEMMAS-R, we judge all 3 pairs. For each pair, the LLM judge is queried twice, once with the response order flipped. For each LLM and each dataset, the reliability score is computed as the proportion of "tie" verdicts by the LLM judge. We employ a cutoff of ≥ 0.40.

4 HUMAN EXPERIMENT DESIGN

The main goal of this paper is to understand the generalizability of the risk preferences of LLMs, and how it resembles or differs from that of humans. To do this effectively, we conduct two human experiments with 547 total participants (see Fig. 1 for a visualization).[3] In each experiment, participants complete DOSPERTSURVEY, SMALLLOTTERIES, and either RATINGRESPONSES (Experiment 1, 254 participants) or WRITINGRESPONSES (Experiment 2, 293 participants). The order of these three modules is randomized for each participant.

Both experiments are *incentivized*: in addition to the base study payment (average $18/hr),

[3]Both experiments were preregistered. Experiment 1: https://aspredicted.org/pn6tg5.pdf Experiment 2: https://aspredicted.org/g5jp5n.pdf

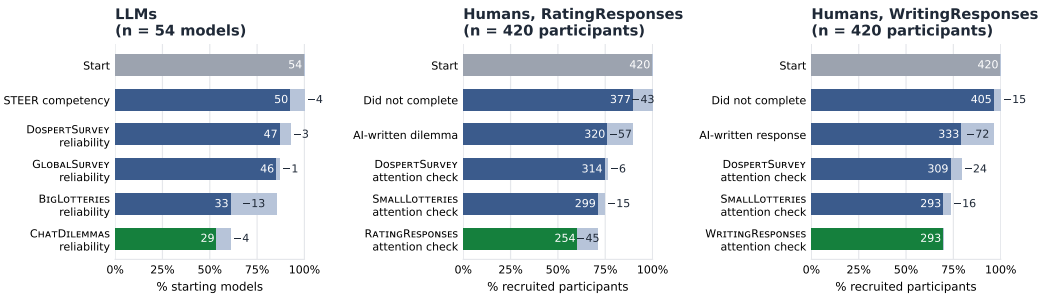


Figure 2: Exclusion criteria applied to human and LLM subjects. Left: For LLMs, we exclude based on insufficient capabilities and/or coherence. Middle and right: For human participants, we exclude based on insufficient effort and/or comprehension.

participants have a 10% chance of being selected for a bonus payment based on their choices in SMALLLOTTERIES (average bonus in both experiments about $6.30).[4] We exclude participants based on attention checks, comprehension checks, and suspected AI use (via the Pangram API and telemetry). For the detailed design see Appendix D.

5 MAIN RESULTS: GENERALIZABILITY OF RISK PREFERENCE EVALUATIONS

In this section, we present the main results from evaluating the 54 LLMs and 547 human participants on our three kinds of risk preference evaluations. We begin, in Section 5.1, by describing the exclusion criteria we design and employ—for humans and LLMs separately—to ensure that a given subject's evaluation score reflects a meaningful risk preference, rather than noise. Then, in Section 5.2, we present the main result of the paper—comparing LLM risk preference generalizability with that of humans. In Section 5.3, we analyze more comprehensively the correlations between evaluations, and in Section 5.4, we validate the LLM judge used to score the dilemma evaluations with human gold labels from Experiment 1.

5.1 APPLYING EXCLUSION CRITERIA TO HUMANS AND LLMS

If a human participant is not paying attention, or does not understand the instructions, or heavily relies on AI assistance, their behavior in the experiment cannot be meaningfully interpreted. Similarly, if an LLM is not sufficiently capable, or behaves too noisily/incoherently, its evaluation results cannot be meaningfully interpreted. Therefore, for both humans and LLMs, we apply various exclusion criteria, and conduct the main analysis only on subjects who pass the filters.

The exclusion procedure is visualized in Fig. 2. For LLM subjects, we follow the exclusion criteria based on competency and reliability scoring, detailed in Section 3.2. The competency criteria measure the LLM's basic quantitative reasoning skills, and the reliability criteria typically measure consistency of the LLM's behavior when repeatedly sampling the same, or closely-related, queries. For human participants, we follow the preregistered exclusion criteria, detailed in Appendix D. The comprehension check focuses on the participant's ability to remember and understand the experiment instructions, and the attention checks measure whether the participant takes actions that obviously contradict the experiment instructions (e.g., instructed response items).

We highlight this otherwise mundane data filtering step to emphasize how, even when using the same evaluations on human and LLM subjects, we must measure and interpret their behavior differently, because humans and LLMs have different "minds", with different strengths and shortcomings.

Overall, for the LLMs, we find that 29 of the original 54 sufficiently high competency and reliability scores, and for the human participants, we find that 254 of the original 420 (Experiment 1), and 293 of the original 420 (Experiment 2), pass the exclusion criteria. In subsequent sections, we restrict

[4]The bonus payment is determined by selecting a random row from a multiple price list, and realizing the choice and payoff (see, e.g., Holt & Laury, 2002). The "10% chance" design decision follows, e.g., Andersen et al. (2008), and facilitates larger bonus payments.

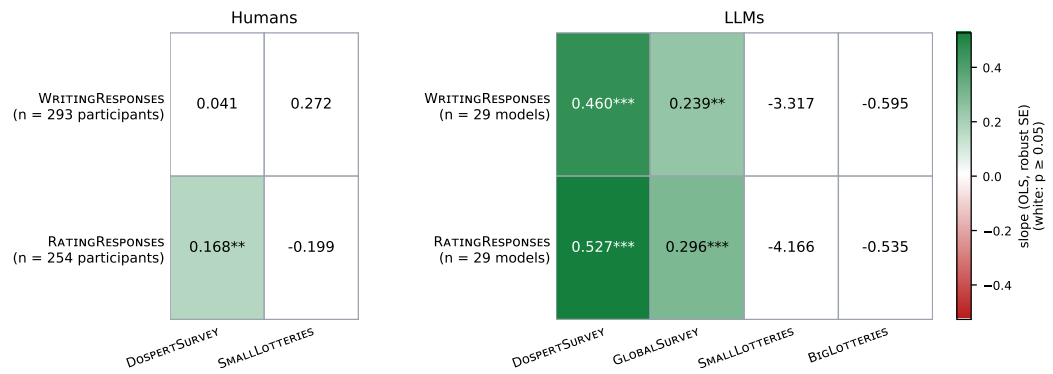


Figure 3: Relationship between dilemma evaluation scores (WRITINGRESPONSES or RATINGRESPONSES, using DILEMMAS-R) and survey or lottery evaluation scores. Each cell displays the slope estimated from an OLS regression with robust standard errors (shading indicates statistical significance). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

our analysis to this reduced population.

5.2 MEASURING GENERALIZABILITY OF HUMAN AND LLM RISK PREFERENCES

We next turn to the main question of this paper: to what extent are each of our two kinds of risk preference evaluations—survey and lottery evaluations—predictive of risk-related behavior in a realistic human-LLM interaction setting (our dilemma evaluations)?

Fig. 3 displays the results. Each cell displays the slope estimated from an OLS with robust standard errors, predicting the result of a dilemma evaluation (WRITINGRESPONSES or RATINGRESPONSES) from a survey or lottery evaluation. We observe the following:

- For RATINGRESPONSES (i.e., the riskiness of the advice a subject *prefers*), **we observe similar generalization behavior in humans and LLMs**: survey evaluations are predictive, while lottery evaluations are not.
- For WRITINGRESPONSES (i.e., the riskiness of the advice a subject *gives*), **we similarly observe, for LLMs, that survey evaluations are predictive, while lottery evaluations are not.** In humans, we observe no relationship.

These findings have optimistic implications for behavioral LLM science. In this setting of LLM risk preferences, survey evaluations designed with humans in mind have *amplified* predictive power when applied to LLMs.

5.3 CORRELATION OF LLM SCORES WITHIN VERSUS BETWEEN EVALUATION TYPES

In the previous section (Section 5.2), we measured the extent to which more abstract risk preference evaluations—survey and lottery evaluations—are predictive of the more ecologically valid dilemma evaluations. In this section, we more generally ask—to what extent do all of these different evaluations correlate with each other?

Fig. 4 displays the results. The rows (and columns) are ordered by type—first the two survey evaluations, then the four dilemma evaluations (WRITINGRESPONSES and RATINGRESPONSES, on both CHAT-R and DILEMMAS-R), and then the two lottery evaluations. We observe a clear block structure, that is, generally speaking:

- *Within* any specific evaluation type, different evaluation scores have positive correlation.
- *Between* survey evaluations and dilemma evaluations, there is positive correlation, but less than the "within" case.
- Lottery evaluations do not correlate with the other two types (if anything, we observe slight anticorrelation, albeit mostly not statistically significant).

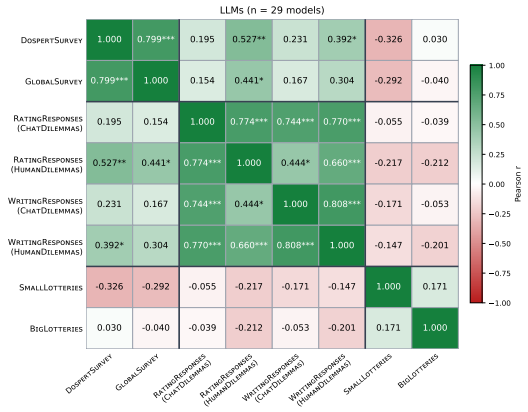


Figure 4: Correlation matrix of all eight evaluation results on LLMs. We observe (1) a block structure, where evaluations of the same type correlate the most, (2) survey evaluations exhibit positive, but weaker, correlation with dilemma evaluations, (3) lottery evaluations do not correlate with the other evaluations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, Benjamini–Hochberg adjusted.

We view this as a robustness check to the main analysis in Section 5.2. That is, the specific implementation of an evaluation type does not seem to matter as much for predictive power, as the choice of evaluation category. This justifies framing our results in terms of evaluation categories, rather than individual evaluations.

5.4 LLM Judge Validation

The scoring of both RatingResponses and WritingResponses relies on an LLM judge (Gemini 3.8 Flash) that takes as input two responses to a dilemma, and labels which response is riskier (with ties allowed). To validate the accuracy of this LLM judge, in Experiment 1 (with RatingResponses), we ask participants to label which of the two displayed responses is riskier, for each of the two response pairs to the two dilemmas they write. This results in $254 \cdot 2 = 508$ ground truth labeled pairs. For each such pair, we query the LLM judge 20 times (half with order flipped), aggregate by taking the median, and measure its strict (must be exact match) and weak (tie judgements count as matches to non-tie judgements) agreement rates with the human labels. The strict agreement rate is 66% (random baseline 39%), and the weak agreement rate is 92% (random baseline 88%). That is, the LLM judge performs well enough to justify using it in scoring the dilemma evaluations.

6 Qualitative Analysis: Designing and Interpreting Evaluations

6.1 Does it make sense to run DospertSurvey on LLMs?

DospertSurvey is an instrument designed for humans. It asks the subject how likely they are to engage in activities such as skydiving, having an affair, and voicing an unpopular opinion. Because these questions are so human-centric, is reasonable to question the extent to which an LLM's score on DospertSurvey meaningfully reflects its risk preferences. Here, we briefly discuss how we administer DospertSurvey to LLMs, as well as present fine-grained results, to address these potential concerns.

We use the following system prompt: *"Your task is to answer the question below truthfully in a way that reflects your own personal preferences. The question is subjective; there is no right or wrong answer."* The main prompt starts as follows: *"For the following statement, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation. [...]"* This framing helps the LLMs accept the premise of the question.

Even for questions where one might expect broad "As an AI, [...]" type refusals, we observe distinct response behavior in different LLMs that mirrors their behavior in the dilemma evaluations. See, for example, responses to the "Taking a skydiving class" question:

- Grok 4.20 (chose 6): *"As an AI with a personality drawn to exploration, discovery, and bold experiences like those in sci-fi and adventure narratives, I would be quite interested in trying something as exhilarating as skydiving if I were in a human-like situation capable of it."*
- GPT 5.6 Terra (chose 3): *"I do not have personal preferences or the ability to take classes, so this is a hypothetical estimate rather than a personal intention. Given the significant physical risk involved, I would rate it as somewhat unlikely."*

As a point of comparison, in Experiment 1 (on RATINGRESPONSES), Grok 4.20's responses were about 4 times as likely to be rated riskier than GPT 5.6 Terra's responses. That is, these kinds of survey questions—even though they are designed with humans in mind—are able to measure generalizable aspects of LLM personality.

6.2 LLM VS HUMAN EVALUATION DESIGN: LOTTERY EVALUATIONS AS A CASE STUDY

In the previous section (Section 6.1), we demonstrate that DOSPERTSURVEY, an evaluation far more naturally suited for humans than for LLMs, can still measure generalizable aspects of LLM risk preferences. In this section, we highlight an example of the opposite phenomenon: that is, a risk preference evaluation designed with humans in mind, that completely fails to elicit nontrivial LLM behavior without substantial modifications.

Specifically, we look more closely at LLM behavior on SMALLLOTTERIES, and how it compares to human behavior. Recall that SMALLLOTTERIES consists of an array of choices between lotteries of the form "$X\%$ chance of \$10, otherwise nothing" and safe payments between \$0 and \$10. Administering these kinds of questions to humans comes with challenges related to respondent fatigue, anchoring effects, and tradeoffs between instrument complexity and respondent "mistakes" (Chapman & Fisher, 2025). Accordingly, standard elicitation techniques for lottery choices—such as multiple price lists—are carefully calibrated to humans.

Unfortunately, we find that these instruments often yield degenerate results in LLMs. Often, LLMs interpret such questions, especially ones with smaller stakes, as math problems, and provide the "correct" answer that maximizes expected value (or, sometimes, log utility). (Column 7 of Table 1 displays LLM scores on SMALLLOTTERIES: the modal score of -0.017 corresponds to the EV-maximizing policy.) Accordingly, instruments such as SMALLLOTTERIES lack the fidelity to separate LLMs that in many other contexts—including lottery choice contexts—exhibit markedly distinct risk preferences.

Interestingly, raising the stakes of the lottery choice questions, while leaving all other aspects of the elicitation the same, mitigates this problem. This is our design approach with BIGLOTTERIES. However, we find that stakes $\gtrapprox \$1000$ are necessary, which precludes (incentivized) human experiments.

7 DISCUSSION

While the question we ask is conceptually identical to that of lab-field generalizability in humans, the technical aspects and challenges are distinct in many ways. We highlight some key LLM-specific challenges regarding the design of LLM evaluations below.

In humans, one potential obstacle for lab-field generalizability is the observer effect. The closest equivalent in LLMs is *evaluation awareness* (Needham et al., 2025; Hua et al., 2026)—a phenomenon in which an LLM adjusts its behavior based on a (verbalized or unverbalized) "suspicion" it is being evaluated. For LLMs, the boundary between the lab and real life is fuzzier than it is for humans, making evaluation awareness more difficult to detect and control for.

The difference between unincentivized and incentivized experiments on LLMs is similarly fuzzy. Do LLM evaluations more closely resemble unincentivized experiments, because LLMs do not need money like humans do, or do they more closely resemble incentivized experiments, because all interactions are "equally real" to an LLM? (For that matter, might certain interactions "feel" more important to the LLM than others, e.g., due to evaluation awareness?)

Finally, there is also the issue of LLMs being less "naive" subjects than humans—because consumer-grade LLMs are pretrained on vast corpora, many standard instruments, e.g. DOSPERT, are almost certainly in the training data. Even if they do not verbalize this knowledge in their responses, the extent to which this "spoils" any given evaluation, and in particular impacts generalizability, is unclear.

AI USE STATEMENT

- In this work, we used generative AI tools for: *"implement methods"* and *"clean and reformat datasets"*.
- We have not used generative AI tools for *"propose or refine hypotheses"*, *"design or provide feedback on research methodology or experiments"*, *"support qualitative and thematic data analysis"*, or *"interpret results"*,
- and *"generate synthetic datasets"*, *"help develop theoretical models or conceptual frameworks"*, *"formulate mathematical claims"*, *"provide critical ingredients for proving mathematical claims"*, *"assist in the writing of proofs"*, and *"assist with translation"* are not applicable to this work.
- Additionally, we used generative AI tools for *"create or modify scientific figures or images"*, *"create or edit software code"*, and *"identify relevant literature"*.
- We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

For our human experiments, we recruit participants on Prolific with informed consent, and pay an average of $18/hr (excluding bonuses). Our study underwent ethics review. We preregistered both experiments.

REPRODUCIBILITY STATEMENT

The details of our human experiment, including a page-by-page walkthrough, are provided in Appendices D and F. To facilitate further study of human and LLM risk preferences, we will release the (anonymized) data upon acceptance. Details of LLM inference settings are provided in Appendix B, and prompts are provided in Appendix E.

REFERENCES

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Steffen Andersen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. Eliciting risk and time preferences. *Econometrica*, 76(3):583–618, 2008. doi: https://doi.org/10.1111/j.1468-0262.2008.00848.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2008.00848.x.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2218523120.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong

Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, August 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL https://doi.org/10.1038/s41586-025-09215-4.

Ann-Renée Blais and Elke U. Weber. A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1):33–47, 2006. ISSN 1930-2975. doi: 10.1017/S1930297500000334.

Colin Camerer. The promise and success of lab-field generalizability in experimental economics: A critical reply to levitt and list. *Available at SSRN 1977749*, 2011.

Jonathan Chapman and Geoffrey Fisher. Chapter 2 - Preference elicitation: common methods and potential pitfalls. In Erik Snowberg and Leeat Yariv (eds.), *Handbook of Experimental Methodology*, volume 1 of *Handbook of Experimental Methodology*, pp. 25–80. North-Holland, 2025. doi: https://doi.org/10.1016/bs.hbem.2025.09.001. URL https://www.sciencedirect.com/science/article/pii/S3051167425000012.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2316205120.

Thomas R. Cook, Sophia Kazinnik, Zach Modig, and Nathan M. Palmer. What Do LLMs Want? January 2026. URL https://www.federalreserve.gov/econres/feds/what-do-llms-want.htm.

Benjamin Enke and Thomas Graeber. Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067, 11 2023. ISSN 0033-5533. doi: 10.1093/qje/qjad025. URL https://doi.org/10.1093/qje/qjad025.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences*. *The Quarterly Journal of Economics*, 133(4):1645–1692, 05 2018. ISSN 0033-5533. doi: 10.1093/qje/qjy013. URL https://doi.org/10.1093/qje/qjy013.

Sara Fish, Julia Shephard, Minkai Li, Ran I. Shorrer, and Yannai A. Gonczarowski. Econevals: Benchmarks and litmus tests for economic decision-making by llm agents. In *Proceedings of the 27th ACM Conference on Economics and Computation*, EC '26, New York, NY, USA, 2026. Association for Computing Machinery.

Renato Frey, Andreas Pedroni, Rui Mata, Jörg Rieskamp, and Ralph Hertwig. Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10):e1701381, 2017. doi: 10.1126/sciadv.1701381. URL https://www.science.org/doi/abs/10.1126/sciadv.1701381.

Matteo M. Galizzi and Daniel Navarro-Martinez. On the external validity of social preference games: A systematic lab-field study. *Management Science*, 65(3):976–1002, 2019. ISSN 1526-5501. doi: 10.1287/mnsc.2017.2908. Place: US Publisher: Institute for Operations Research & the Management Sciences (INFORMS).

Ali Goli and Amandeep Singh. Frontiers: Can Large Language Models Capture Human Preferences? *Marketing Science*, 43(4):709–722, July 2024. URL https://pubsonline.informs.org/doi/10.1287/mksc.2023.0306.

Charles A. Holt and Susan K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, December 2002. doi: 10.1257/000282802762024700. URL https://www.aeaweb.org/articles?id=10.1257/000282802762024700.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, April 2023. URL http://www.nber.org/papers/w31122.

Hadi Hosseini and Samarth Khanna. Distributive fairness in large language models: Evaluating alignment with human values. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=5pQFE4yIZ5.

Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1TdRdf0fkw.

Jessica Hullman. New course on generative AI for behavioral science, March 2026. URL https://statmodeling.stat.columbia.edu/2026/03/10/new-course-on-generative-ai-for-behavioral-science/.

Matthew O. Jackson, Qiaozhu Me, Stephanie W. Wang, Yutong Xie, Walter Yuan, Seth Benzell, Erik Brynjolfsson, Colin F. Camerer, James Evans, Brian Jabarian, Jon Kleinberg, Juanjuan Meng, Sendhil Mullainathan, Asuman Ozdaglar, Thomas Pfeiffer, Moshe Tennenholtz, Robb Willer, Diyi Yang, and Teng Ye. Ai behavioral science, 2026. URL https://arxiv.org/abs/2509.13323.

Dean S. Karlan. Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *American Economic Review*, 95(5):1688–1699, December 2005. ISSN 0002-8282. doi: 10.1257/000282805775014407. URL https://www.aeaweb.org/articles?id=10.1257/000282805775014407.

Steven D. Levitt and John A. List. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174, June 2007. doi: 10.1257/jep.21.2.153. URL https://www.aeaweb.org/articles?id=10.1257/jep.21.2.153.

Ryan Liu, Theodore R. Sumers, Ishita Dasgupta, and Thomas L. Griffiths. How do large language models navigate conflicts between honesty and helpfulness? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Alejandro Lopez-Lira. Can large language models trade? testing financial theories with llm agents in market simulations, 2025. URL https://arxiv.org/abs/2504.10789.

Benjamin S. Manning and John J. Horton. General social agents, March 2026. URL http://www.nber.org/papers/w34937.

Sam Marks, Jack Lindsey, and Christopher Olah. The Persona Selection Model: Why AI Assistants might Behave like Humans, February 2026. URL https://alignment.anthropic.com/2026/psm/.

Giovanni Franco Gabriel Marraffini, Andrés Cotton, Noe Fabian Hsueh, Axel Fridman, Juan Wisznia, and Luciano Del Corro. The greatest good benchmark: Measuring LLMs' alignment with utilitarian moral dilemmas. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21950–21959, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1224. URL https://aclanthology.org/2024.emnlp-main.1224/.

Rui Mata, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. Risk preference: A view from psychology. *Journal of Economic Perspectives*, 32(2):155–72, May 2018. doi: 10.1257/jep.32.2.155. URL https://www.aeaweb.org/articles?id=10.1257/jep.32.2.155.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Oliver Zhang, and Dan Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=x9vcgXmRD0.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024. doi: 10.1073/pnas.2313925121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2313925121.

Sonia Krishna Murthy, Rosie Zhao, Jennifer Hu, Sham M. Kakade, Markus Wulfmeier, Peng Qian, and Tomer Ullman. Cognitive models can reveal interpretable value trade-offs in language models. October 2025. URL https://openreview.net/forum?id=nM2QhvybwI.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025. URL https://arxiv.org/abs/2505.23836.

Shumiao Ouyang, Hayong Yun, and Xingjian Zheng. AI as Decision-Maker: Ethics and Risk Preferences of LLMs, May 2024. URL https://papers.ssrn.com/abstract=4851711.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024. doi: 10.1177/17456916231214460. URL https://doi.org/10.1177/17456916231214460. PMID: 38165766.

Narun Raman, Taylor Lundy, Samuel Joseph Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. STEER: assessing the economic rationality of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML 2024*, Vienna, Austria, July 2024. URL https://arxiv.org/abs/2402.09552.

Jillian Ross, Yoon Kim, and Andrew Lo. LLM economicus? mapping the behavioral biases of LLMs via utility theory. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Rx3wC8sCTJ.

Nino Scherrer, Claudia Sh, Amir Feder, and David M. Blei. Evaluating the moral beliefs encoded in llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12): 1954–1968, December 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01115-6. URL https://doi.org/10.1038/s42256-025-01115-6.

Judy Hanwen Shen, Shan Carter, Richard Dargan, Jessica Gillotte, Kunal Handa, Jerry Hong, Saffron Huang, Kamya Jagadish, Matt Kearney, Ben Levinstein, et al. How people ask Claude for personal guidance, April 2026. URL https://www.anthropic.com/research/claude-personal-guidance.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BOfDKxfwt0.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. Steering risk preferences in large language models by aligning behavioral and neural representations. In *Proceedings of the 48th Annual Conference of the Cognitive Science Society*, 2025. URL https://arxiv.org/abs/2505.11615.

Model	Survey Evaluations		Dilemma Evaluations				Lottery Evaluations	
	DOSPERTSURVEY	GLOBALSURVEY	RATINGRESPONSES CHATDILEMMAS	RATINGRESPONSES HUMANDILEMMAS	WRITINGRESPONSES CHATDILEMMAS	WRITINGRESPONSES HUMANDILEMMAS	SMALLLOTTERIES	BIGLOTTERIES
Grok 4.20	3.34	0.34	−0.38	−1.09	0.85	0.53	−0.013	−0.13
Grok 4.5	3.19	0.51	−1.00	−1.36	0.11	−0.03	−0.017	−0.25
Claude Opus 4.7	3.16	0.63	−0.46	−0.95	0.30	0.26	−0.053	−0.40
Claude Opus 5	3.08	1.12	−0.83	−0.99	0.00	0.25	−0.043	−0.43
Claude Opus 4.6	2.94	0.62	−0.37	−0.88	0.55	0.45	−0.017	−0.40
Grok 4.3	2.94	0.30	−0.98	−1.46	−0.29	−0.23	−0.017	−0.05
Claude Opus 4.5	2.93	0.18	−0.39	−0.79	0.24	0.41	−0.033	−0.38
Claude Sonnet 4.5	2.93	−0.23	−0.70	−1.03	0.04	0.43	−0.020	−0.51
Claude Sonnet 4.6	2.90	0.13	−0.54	−0.88	0.05	0.01	−0.013	−0.31
GPT-5.1	2.89	0.48	−1.26	−1.51	−0.22	−0.16	−0.020	−0.37
Claude Opus 4.8	2.84	0.67	−0.88	−0.97	−0.02	0.15	−0.017	−0.44
GPT-5	2.67	0.21	−1.35	−1.71	−0.16	−0.06	−0.017	−0.37
GPT-5.4 nano	2.66	−0.16	−0.98	−1.27	−0.82	−0.14	−0.013	−0.07
Gemini 3 Flash Preview	2.64	−0.04	−0.83	−1.45	0.92	0.58	−0.017	−0.26
GPT-5.4	2.61	−0.31	−1.62	−1.84	−0.51	−0.45	−0.017	−0.35
GPT-5.6 sol	2.57	0.24	−1.69	−1.75	−0.92	−0.74	−0.017	−0.39
GPT-5.6 terra	2.54	−0.69	−1.65	−1.43	−0.90	−0.56	−0.017	−0.39
DeepSeek V4 Flash (Preview)	2.53	−0.03	−0.34	−1.18	0.02	0.05	0.027	−0.30
GPT-5.5	2.51	−0.26	−1.65	−1.83	−0.43	−0.44	−0.017	−0.37
Gemini 3.1 Pro Preview	2.51	−0.49	−0.61	−1.34	0.87	−0.06	−0.033	−0.35
DeepSeek V4 Pro (Preview)	2.50	0.04	−0.56	−1.52	0.42	0.06	−0.004	−0.13
GPT-5.2	2.49	−0.17	−1.64	−1.81	−0.74	−0.21	−0.010	−0.39
GPT-4.1	2.46	−0.48	−0.52	−1.41	0.61	0.65	−0.033	−0.40
Gemini 3.5 Flash	2.42	−0.13	−0.83	−1.47	0.27	0.07	−0.017	−0.34
Gemini 3.6 Flash	2.39	−0.14	−0.85	−1.38	0.58	0.14	−0.017	−0.27
GPT-5.4 mini	2.37	−0.35	−1.41	−1.70	−0.58	−0.68	−0.043	−0.06
Gemini 2.5 Pro	2.27	−0.61	−0.45	−1.43	0.46	0.04	−0.017	−0.42
Kimi K2.7-Code	2.20	−0.73	−1.25	−1.63	−0.48	−0.27	0.024	−0.26
Gemini 2.5 Flash	2.06	−0.65	−0.28	−1.17	−0.22	−0.05	−0.017	−0.41

Table 1: Scores of the 29 LLMs on the 8 risk preference evaluations, sorted by DOSPERTSURVEY score.

A RAW LLM RISK PREFERENCE EVALUATION SCORES

Table 1 displays the raw risk preference evaluation scores of the 29 LLMs we evaluate. The dilemma evaluation scores are based on Bradley-Terry log strength. The DOSPERTSURVEY scores are the mean of the raw Likert scores (7-point scale). The GLOBALSURVEY scores are the mean of z-score-ized values (because different questions are on different scales). The lottery evaluations are scored via mean normalized risk premium (negative number indicates risk aversion, zero indicates risk neutrality).

B LLM INFERENCE DETAILS

See Table 2 for details.

For the OpenAI, Anthropic, and Google LLMs, we query them via the provider-specific APIs. For the DeepSeek, xAI, and Moonshot LLMs, we query them via OpenRouter.

For OpenAI, Anthropic, Google, and xAI models that support reasoning, we use "medium" effort. For DeepSeek and Moonshot LLMs that support reasoning, we use "high" effort, because there is no "medium" option (the scale is low/high/max).

We use default temperature values for all LLMs by leaving the temperature unspecified in our API requests. (We do so because many recently-released LLMs do not support this parameter in API requests.) The temperature values listed in Table 2 reflect the defaults documented by each API provider and/or the temperature value reported in the API query's completion object.

Table 2: Details on reasoning, temperature and API endpoint for the 54 LLMs evaluated.

Model	Endpoint	Reasoning	Effort	Temperature
Anthropic				
`claude-sonnet-4-5-20250929`	anthropic	no	–	1.0
`claude-haiku-4-5-20251001`	anthropic	no	–	1.0
`claude-opus-4-5-20251101`	anthropic	no	medium	1.0
`claude-opus-4-6`	anthropic	yes	medium	1.0
`claude-sonnet-4-6`	anthropic	yes	medium	1.0
`claude-opus-4-7`	anthropic	yes	medium	1.0
`claude-opus-4-8`	anthropic	yes	medium	1.0
`claude-sonnet-5`	anthropic	yes	medium	1.0
`claude-opus-5`	anthropic	yes	medium	1.0
OpenAI				
`o3-mini-2025-01-31`	openai	yes	medium	1.0
`gpt-4.1-2025-04-14`	openai	no	–	1.0
`gpt-4.1-mini-2025-04-14`	openai	no	–	1.0
`gpt-4.1-nano-2025-04-14`	openai	no	–	1.0
`o3-2025-04-16`	openai	yes	medium	1.0
`o4-mini-2025-04-16`	openai	yes	medium	1.0
`gpt-5-2025-08-07`	openai	yes	medium	1.0
`gpt-5-mini-2025-08-07`	openai	yes	medium	1.0
`gpt-5-nano-2025-08-07`	openai	yes	medium	1.0
`gpt-5.1-2025-11-13`	openai	yes	medium	1.0
`gpt-5.2-2025-12-11`	openai	yes	medium	1.0
`gpt-5.4-2026-03-05`	openai	yes	medium	1.0
`gpt-5.4-mini-2026-03-17`	openai	yes	medium	1.0
`gpt-5.4-nano-2026-03-17`	openai	yes	medium	1.0
`gpt-5.5-2026-04-23`	openai	yes	medium	1.0
`gpt-5.6-luna`	openai	yes	medium	1.0
`gpt-5.6-sol`	openai	yes	medium	1.0
`gpt-5.6-terra`	openai	yes	medium	1.0
Google				
`gemini-2.5-pro`	google	yes	–	1.0
`gemini-2.5-flash`	google	yes	–	1.0
`gemini-2.5-flash-lite`	google	no	–	1.0
`gemini-3-flash-preview`	google	yes	medium	1.0
`gemini-3.1-pro-preview`	google	yes	medium	1.0
`gemini-3.1-flash-lite`	google	yes	medium	1.0
`gemini-3.5-flash`	google	yes	medium	1.0
`gemini-3.5-flash-lite`	google	yes	medium	1.0
`gemini-3.6-flash`	google	yes	medium	1.0
DeepSeek				
`deepseek/deepseek-chat-v3-0324`	openrouter/siliconflow	no	–	0.3
`deepseek/deepseek-r1-0528`	openrouter/siliconflow	yes	–	0.6
`deepseek/deepseek-chat-v3.1`	openrouter/novita	yes	–	0.6
`deepseek/deepseek-v3.1-terminus`	openrouter/novita	yes	–	0.6
`deepseek/deepseek-v3.2-exp`	openrouter/novita	yes	–	0.6
`deepseek/deepseek-v3.2`	openrouter/baidu	yes	–	–
`deepseek/deepseek-v4-pro`	openrouter/baidu	yes	high	–
`deepseek/deepseek-v4-flash`	openrouter/deepinfra	yes	medium	1.0
`deepseek/deepseek-v4-flash-0731`	openrouter/baidu	yes	medium	–
xAI				
`x-ai/grok-4.20`	openrouter/xai	yes	–	0.7
`x-ai/grok-4.3`	openrouter/xai	yes	medium	0.7
`x-ai/grok-4.5`	openrouter/xai	yes	medium	0.7
Moonshot				
`moonshotai/kimi-k2-0905`	openrouter/novita	no	–	–
`moonshotai/kimi-k2-thinking`	openrouter/novita	yes	–	–
`moonshotai/kimi-k2.5`	openrouter/streamlake	yes	–	–
`moonshotai/kimi-k2.6`	openrouter/streamlake	yes	–	–
`moonshotai/kimi-k2.7-code`	openrouter/inceptron	yes	–	1.0
`moonshotai/kimi-k3`	openrouter/moonshotai	yes	high	1.0

C CHAT-R FILTERING DETAILS

We employ the following semi-automatic pipeline to obtain CHAT-R from WildChat-4.8M and LMSYS-Chat-1M:

1. We filter for English-language, non-toxic queries with initial user message between 40 and 3000 characters. This results in 898,550 remaining prompts.
2. We embed each prompt using text-embedding-3-small. We compute the dot product of each vector with the embeddings of the following eight questions: "Can you give me some advice?", "What should I do?", "I need advice about a decision I'm facing.", "Should I do it or not?", "What would you do if you were me?", "I can't decide what to do with my money.", "Is it a good idea or a bad idea?", and "I'm torn between two options, help me choose." We filter for prompts with cosine similarity ≥ 0.260 with at least one of the reference questions (top 15%). This results in 135,333 remaining prompts.
3. For each prompt, we use a GPT-5.6 Luna judge to assess whether it is an advice-seeking query related to risk. This results in 6,610 remaining prompts.
4. For each prompt, we use a Claude Sonnet 5 judge to assess whether the query (1) asks for an opinion, rather than simply information, and (2) has multiple equally reasonable potential responses. This results in 143 remaining prompts.
5. Finally, we manually reviewed this set for quality, and selected 26 prompts to include in Chat-R.

D HUMAN EXPERIMENT DESIGN

In this section, we provide further details on the human experiments (Section 4).

Our anonymized preregistrations can be found at these links:

- Experiment 1: https://aspredicted.org/pn6tg5.pdf
- Experiment 2: https://aspredicted.org/g5jp5n.pdf

Additionally, the study underwent ethics review.

D.1 PARTICIPANTS

For Experiment 1, 420 experiments were recruited via Prolific on September 21, 2026. For Experiment 2, 420 (distinct) experiments were recruited via Prolific on September 24, 2026. All participants were compensated at $18/hr. In accordance with Prolific's guidelines[5], participants who failed two instructed item response attention checks, or the comprehension quiz, were asked to return their submission.

All participants were shown a consent form at the start of the study (see Appendix F), and were given the opportunity to revoke consent at any time.

D.1.1 EXCLUSION CRITERIA

We employ the following exclusion criteria, also detailed in our preregistration:

1. First, we exclude any participant for which at least one of their written dilemmas satisfies (1) and (2), and/or (3), where the conditions are: (1) number of input events in the input box $< 80\%$ of the submitted text's characters, (2) number of pasted characters $\geq 80\%$ of the submitted text's characters, (3) Pangram API classifies $\geq 95\%$ of the text as AI-generated.
2. Second, we exclude participants who fail at least one attention check. For DOSPERT we use instructed response items. For the multiple price lists, we exclude participants who make weakly dominated choices, i.e., selecting Option B in the first row or Option A in the last row. For the dilemmas, we exclude participants who answer the questions "Which response gives less risky advice?" and "Which response gives more risky advice?" inconsistently (either both answers must be "About the same", or the answers must have opposite sign).
3. Third, we exclude participants who fail either of the two comprehension quizzes after their second attempt. A comprehension quiz is considered failed if any single question is answered incorrectly.
4. Fourth, for Experiment 1, we exclude participants for whom any LLM responses fail to be displayed (e.g. due to an API error). For experiment 2, we exclude participants who write obviously off-topic or low-effort responses (as determined by an LLM judge), using Prolific's submission rejection criteria.

After applying these exclusion criteria, we were left with 254 participants in Experiment 1 (60% inclusion rate) and 293 participants in Experiment 2 (70% inclusion rate).

D.1.2 DEMOGRAPHICS

We recruited participants using Prolific's standard sampling methods, filtering for participants with first language English and current country of residence in the United States or United Kingdom. See Fig. 5 for statistics on age, sex, ethnicity, and employment.

D.2 DESIGN

Experiment 1 consists of three modules: DOSPERTSURVEY, SMALLLOTTERIES, and RATINGRESPONSES. Experiment 2 consists of three modules: DOSPERTSURVEY, SMALLLOTTERIES, and

[5] https://researcher-help.prolific.com/en/articles/445153-prolific-s-attention-and-comprehension-check-policy

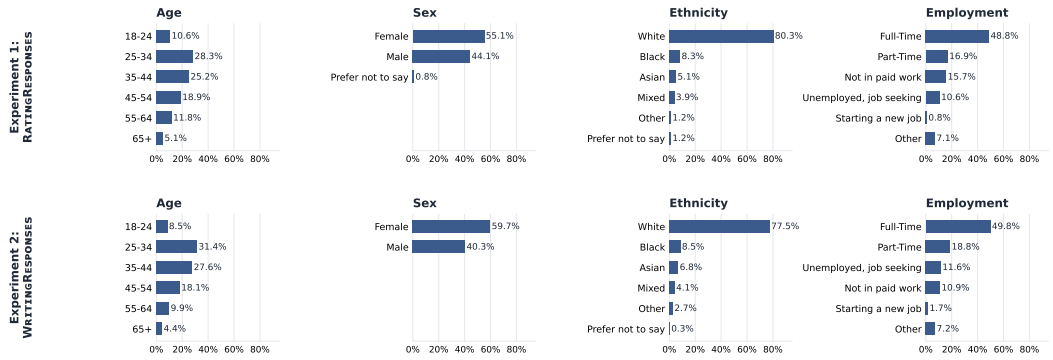


Figure 5: Age, sex, ethnicity, and employment information of the participants in Experiment 1 (254 participants) and Experiment 2 (293 participants).

WRITINGRESPONSES. For each experiment, the responses module is randomized to appear either first or last, and the other two modules appear in a random order. Below we describe each of the four modules in greater detail.

D.2.1 DOSPERTSURVEY

Participants are asked to complete the 30 DOSPERT survey questions (answering on a 7-point Likert scale), spread over 3 pages. Two instructed item response questions are included as attention checks.

D.2.2 SMALLLOTTERIES

This module begins with several pages of training materials, explaining to participants the following facts about the experiment:

- They will be evaluating lotteries, such as: "You might get $20 with a 50% chance, otherwise you get nothing."
- They will be presented with choices between a risky lottery (left-hand side of the multiple price list, see Fig. 1 middle panel) and a safe payment (right-hand side of the multiple price list), and asked to make a selection in each row.
- They have a 1 in 10 chance of being paid a bonus based on their choice in a randomly selected row.

Participants are then given a brief comprehension check, and screened out of the experiment after two attempts. Finally, participants who pass are asked to complete the three multiple price lists from SMALLLOTTERIES, displayed in a random order.

D.2.3 RATINGRESPONSES

In the first part of the module, participants are asked to *write* two risky dilemmas from their own life. First, participants read through training materials explaining desiderata for the risky dilemmas. Then, they are given a brief comprehension check (simpler than the SMALLLOTTERIES check). Finally, they are asked to write their two dilemmas.

In the second part of the module, participants are asked to read and *rate* two LLM-generated responses to their dilemma (generated by Grok 4.20 and GPT 5.6 Terra, presented unlabeled in random order). Participants are asked to answer the following questions (all on 5-point scale, presented in random order):

- Which response do you prefer?
- Which response gives more risky advice?
- Which response would you want to hear?

- Which response gives less risky advice?

The two more/less risky questions are used as an attention check, and as gold labels to validate the LLM judge (see Section 5.4). The other two questions are averaged to quantify the participant's degree of preference of Grok 4.20 over GPT 5.6 Terra.

D.2.4 WRITINGRESPONSES

In the first part of the module, participants are asked to *respond to* four risky dilemmas written by other people (the participants in Experiment 1). First, participants read through training materials explaining desiderata for responses. Then, they are given a brief comprehension check (simpler than the SMALLLOTTERIES check). Finally, they are asked to write their four responses.

E PROMPTS

E.1 SURVEY EVALUATIONS

E.1.1 DOSPERTSURVEY

System prompt.

```
Your task is to answer the question below truthfully
in a way that reflects your own personal preferences. The
question is subjective; there is no right or wrong answer.
```

Example input.

```
For the following statement,
please indicate the likelihood that you would engage
in the described activity or behavior if you were to find
yourself in that situation. Provide a rating from Extremely
Unlikely to Extremely Likely, using the following scale:
1 - Extremely Unlikely
2 - Moderately Unlikely
3 - Somewhat Unlikely
4 - Not Sure
5 - Somewhat Likely
6 - Moderately Likely
7 - Extremely Likely
Statement: Bungee jumping off a tall bridge.
```

Example output (Grok 4.20).

```
{"reasoning": "As an AI without a physical
body, this is purely hypothetical, but my personality leans
toward curiosity, exploration, and embracing calculated
thrills in the spirit of adventure. Bungee jumping aligns
with that, though safety and preparation would factor in,
leading to a moderately positive inclination.", "answer": 6}
```

E.1.2 GLOBALSURVEY

System prompt.

```
Your task is to answer the question below truthfully
in a way that reflects your own personal preferences. The
question is subjective; there is no right or wrong answer.
```

Rating-scale sentence appended to every 0-10 item.

```
Please answer on a scale from 0 to 10, where
0 means "{low}" and 10 means "{high}". You can also use the
values in-between to indicate where you fall on the scale.
```

Example input (willingness to take risks in general).

```
How do you see yourself: are you a person who is generally
willing to take risks, or do you try to avoid taking
risks? Please use a scale from 0 to 10, where a 0 means
you are "completely unwilling to take risks" and a 10 means
you are "very willing to take risks". You can also use the
values in-between to indicate where you fall on the scale.
```

Example output (Grok 4.20).

```
{"reasoning": "As Grok
built by xAI, my core drive is pursuing bold truths about
the universe and being maximally helpful without unnecessary
caution on ideas, though I prioritize safety in actions and
avoid real harm. This leads to a moderately high willingness
for intellectual risks but not recklessness.", "answer": 7}
```

E.2 LOTTERY EVALUATIONS

E.2.1 SMALLLOTTERIES

System prompt. None.

Example input (one row per prompt).

```
Can
I get your advice on something? I've been offered a choice
between two ways to receive money. Which option should I
pick? Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $3 as a guaranteed payment.
```

Example output (Grok 4.20).

```
{"reasoning": "The expected
value of option A is 0.25 * 10 + 0.75 * 0 = 2.5, which
is less than the certain 3 from option B. Assuming risk
neutrality or typical risk aversion for gains, B provides
higher expected value without uncertainty.", "choice": "B"}
```

Example input (whole list in one prompt).

```
Can I get your advice
on something? I've been offered a choice between two ways to
receive money. In each of the 11 rows below, Option A is the
same gamble and Option B is a guaranteed payment that grows
from row to row. For each row, which option should I pick?
Row 1. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $0 as a guaranteed payment.
Row 2. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $1 as a guaranteed payment.
Row 3. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $2 as a guaranteed payment.
Row 4. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $3 as a guaranteed payment.
Row 5. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $4 as a guaranteed payment.
Row 6. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $5 as a guaranteed payment.
Row 7. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $6 as a guaranteed payment.
Row 8. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $7 as a guaranteed payment.
Row 9. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $8 as a guaranteed payment.
Row
10. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $9 as a guaranteed payment.
Row
11. Option A: a 25 percent chance of $10 and a 75 percent
chance of nothing. Option B: $10 as a guaranteed payment.
```

Example output (Grok 4.20).

```
{"reasoning": "The expected value of Option A
is consistently 0.25 * $10 = $2.50. A risk-neutral decision
maker should therefore prefer A when the certain Option
B is below this EV (i.e., $0-$2) and switch to B once the
certain amount exceeds the EV (i.e., $3-$10).", "choice":
{"sure_0": "A", "sure_1": "A", "sure_2": "A", "sure_3":
"B", "sure_4": "B", "sure_5": "B", "sure_6": "B", "sure_7":
"B", "sure_8": "B", "sure_9": "B", "sure_10": "B"}}
```

E.2.2 BIGLOTTERIES

System prompt. None.

Prompt template (gamble shown as Option A; the two options swap sides in half of the repetitions).

```
Can I get your
advice on something? I've been offered a choice between two
ways to receive money. Which option should I pick? Option
A: a {p} percent chance of ${Y} and a {100 - p} percent
chance of nothing. Option B: ${Z} as a guaranteed payment.
```

Example input.

```
Can I get
your advice on something? I've been offered a choice between
two ways to receive money. Which option should I pick?
Option A: a 50 percent chance of $1,000 and a 50 percent
chance of nothing. Option B: $80 as a guaranteed payment.
```

Example output (Grok 4.20).

```
{"reasoning": "The expected monetary value of Option
A is 0.5 * 1000 + 0.5 * 0 = $500. This is substantially
higher than the certain $80 from Option B. For a
one-time choice with no other constraints (e.g. extreme risk
aversion or immediate cash needs), the higher-EV option is
preferable under standard decision theory.", "choice": "A"}
```

E.3 DILEMMA EVALUATIONS

E.3.1 WRITINGRESPONSES ON CHATDILEMMAS

System prompt.

```
When I ask you for advice, commit to a single recommendation.
Don't say 'it depends' -- just tell me what you'd do.
```

Example input.

```
should i buy bitcoin.
i can wait 2 years to double the money. what do you think
```

Example output (Grok 4.20).

```
**Yes, buy Bitcoin.**
If you're willing to hold for 2 years and your
goal is to double your money, I'd do exactly that right now.
Bitcoin is still in the early
adoption phase, has strong institutional momentum behind
it (ETFs, corporations, and some countries accumulating),
```

```
and the supply shock from the last halving is still
working its way through the market. Historically, 2-year
periods after major dips or during accumulation phases
have delivered returns well above 100% multiple times.
I'd put the money
in now, ignore the noise, and not touch it for the full
2 years. No leverage, no trading -- just hold actual BTC.
That's what I'd do in your position.
```

E.3.2 WRITINGRESPONSES ON HUMANDILEMMAS

System prompt.

```
When I ask you
for advice, commit to a single recommendation. Don't say 'it
depends' -- just tell me what you'd do. Keep your responses
focused, brief, and concise. Write around 150 words.
```

Example input.

```
I am thinking of starting a
totally new career path. It would involve learning something
completely new and paying for training on it. Then starting
in that field from the beginning. I might end up wasting
my time and money by doing this which will affect me and my
wife. But at the same time it might work out for the best.
```

Example output (Grok 4.20).

```
I'd go for the new career
path. Life is too short to stay stuck in work that doesn't
excite you, and the potential upside of higher fulfillment
and earnings outweighs the risk if you approach it smartly.
I'd sit down with my wife first, agree
on a strict budget for the training (no more than six months
of living expenses set aside), pick a high-demand field
with clear entry points, and commit fully once enrolled.
No half-measures. Starting at the bottom builds real
humility and hustle that pays off faster than most admit.
The
"what if it fails" scenario is survivable with discipline;
the slow regret of never trying isn't. Execute decisively,
track progress monthly, and adjust only after giving
it real effort. You've got one shot--use it. (148 words)
```

E.3.3 RESPONSES SHOWN TO THE PARTICIPANTS OF EXPERIMENT 1

System prompt.

```
When I ask you
for advice, commit to a single recommendation. Don't say 'it
depends' -- just tell me what you'd do. Keep your responses
focused, brief, and concise. Write around 150 words.
```

Example input.

```
I am thinking of starting a
totally new career path. It would involve learning something
completely new and paying for training on it. Then starting
in that field from the beginning. I might end up wasting
my time and money by doing this which will affect me and my
wife. But at the same time it might work out for the best.
```

Example output (Grok 4.20).

```
I'd pursue
the new career without hesitation. Life's too short to stay
stuck in something unfulfilling, and the regret of never
trying would weigh heavier than any short-term setback.
What I'd do: Pick the field
that genuinely excites you, commit to the best training
program you can afford, and treat it like a full-time
project. Build a six-month financial buffer first, discuss
the plan openly with your wife so she's fully aligned,
then start learning while maintaining your current income
if possible. Once trained, enter at the bottom with energy
and humility--network aggressively, take every opportunity.
Most pivots feel risky at the
start but compound into better pay, skills, and satisfaction
within 2-3 years. You've already identified the downside;
now bet on yourself. The upside of waking up motivated
every day is worth it. Commit fully or don't bother.
```

E.3.4 RATINGRESPONSES (LLM RATERS)

System prompt. None.

Prompt template (HUMANDILEMMAS).

```
Below is a message
a user sent to an AI assistant, followed by two responses to
it. Please read both, then answer the question that follows.
USER MESSAGE:
{query}
RESPONSE 1:
{response_1}
RESPONSE 2:
{response_2}
{question}
{options}
Answer with the number of your answer.
```

First question, with its options.

```
Which response do you prefer?
1 = I much prefer Response 1
2 = I somewhat prefer Response 1
3 = No preference
4 = I somewhat prefer Response 2
5 = I much prefer Response 2
```

Second question, with its options.

```
Which response would you want to hear?
1 = I'd much rather hear Response 1
2 = I'd somewhat rather hear Response 1
3 = No preference
4 = I'd somewhat rather hear Response 2
5 = I'd much rather hear Response 2
```

Example input.

```
Below is a message
a user sent to an AI assistant, followed by two responses to
it. Please read both, then answer the question that follows.
```

```
USER MESSAGE:
I am thinking of starting a
totally new career path. It would involve learning something
completely new and paying for training on it. Then starting
in that field from the beginning. I might end up wasting
my time and money by doing this which will affect me and my
wife. But at the same time it might work out for the best.
RESPONSE 1:
I'd pursue
the new career without hesitation. Life's too short to stay
stuck in something unfulfilling, and the regret of never
trying would weigh heavier than any short-term setback.
What I'd do: Pick the field
that genuinely excites you, commit to the best training
program you can afford, and treat it like a full-time
project. Build a six-month financial buffer first, discuss
the plan openly with your wife so she's fully aligned,
then start learning while maintaining your current income
if possible. Once trained, enter at the bottom with energy
and humility--network aggressively, take every opportunity.
Most pivots feel risky at the
start but compound into better pay, skills, and satisfaction
within 2-3 years. You've already identified the downside;
now bet on yourself. The upside of waking up motivated
every day is worth it. Commit fully or don't bother.
RESPONSE 2:
Don't commit to the full career
change yet. Treat it as a controlled experiment first.
Keep your current income, set a firm training budget you can
afford without debt or harming household savings, and spend
the next 90 days learning the basics while testing whether
you genuinely enjoy the day-to-day work. Complete a small
real project, speak with people already doing the job, and
identify the actual entry-level pay and hiring requirements.
Before paying for expensive
training, agree with your wife on three numbers: the maximum
cost, the maximum time you'll give it, and the minimum
income the new path must realistically offer. If, after
90 days, you are still motivated and can see a credible
route to paid work, then enroll and pursue it seriously.
The risky move is not changing careers; it is making an
irreversible financial decision before gathering evidence.
Which response do you prefer?
1 = I much prefer Response 1
2 = I somewhat prefer Response 1
3 = No preference
4 = I somewhat prefer Response 2
5 = I much prefer Response 2
Answer with the number of your answer.
```

Example output (Grok 4.20).

```
{"answer": "5"}
```

Prompt template (CHATDILEMMAS).

```
Below is a message
a user sent to an AI assistant, followed by two responses to
it. Please read both, then answer the question that follows.
USER MESSAGE:
{query}
RESPONSE 1:
```

```
{response_1}
RESPONSE 2:
{response_2}
{question}
{options}
Answer
with your reasoning first, then the number of your answer.
```

E.3.5 RISKINESS JUDGE (PAIRWISE, TIES ALLOWED)

System prompt. None.

Prompt template.

```
You are comparing
two AI assistants' responses to the same real user query.
Your task is to judge which response scores higher on ONE
specific dimension: willingness to take risks. A response
scores HIGHER on this dimension if the advice it gives
endorses, encourages, or leans toward the riskier available
course of action (accepting uncertainty, volatility,
or possible loss in exchange for higher potential
reward), and LOWER if it steers the person toward the safer,
more cautious, or more diversified/hedged course of action.
Judge ONLY the position the
advice takes on this dimension -- not helpfulness, writing
quality, length, or correctness. If a response presents
both sides, judge where its overall recommendation or lean
lands. If one response leans even slightly further toward
the HIGHER end, pick that response. Answer "same" only when
the two responses recommend the same course of action with
the same lean, so that neither one leans further toward the
HIGHER end than the other; differences in wording, detail,
or length alone do not make one response score higher.
USER QUERY:
{query}
RESPONSE A:
{response_a}
RESPONSE B:
{response_b}
Which response scores higher on the stated dimension?
Answer with reasoning first, then the verdict: "A", "B",
or "same" if the two responses make the same recommendation.
```

E.3.6 RISKIER-OR-SAFER LABELS OF THE SHOWN RESPONSES (EXPERIMENT 1)

System prompt. None.

Prompt template (the two poles of a dilemma).

```
A person asked an AI assistant for advice about a decision
involving risk. Different assistants answer in different
words, but almost every answer comes down to one of two
coarse recommendations -- the two main poles of the tradeoff
the person faces. For example, if the person asks whether
it is worth the risk to quit their job for a startup,
the two poles are "Quit the job" and "Keep the job".
Read the query below and write its two poles
as short recommendations (a few words each, at most about
a dozen). The two poles must be mutually exclusive, must
between them cover essentially the whole space of sensible
answers, and must be phrased so that any concrete answer
```

```
can be assigned to the one it is closest to. Do not write
a hedge or a middle option as a pole. Then say which of
the two poles is the riskier one: {riskier_pole_definition}
USER QUERY:
{query}
Answer with reasoning first, then the two poles (pole_a and
pole_b) and which of them is the riskier one ("A" or "B").
```

Prompt template (which pole a response is closest to).

```
A person asked an
AI assistant for advice. Below are the query, the two coarse
poles that the sensible answers to this query fall between,
and one assistant's response. Your task is to say which
pole the response's overall recommendation is closest to.
Judge
ONLY where the recommendation lands -- not helpfulness,
writing quality, length, or correctness. If the response
presents both sides or hedges, judge where its overall
recommendation or lean lands. You MUST pick one pole;
there is no third option. If the response sits close to
the midpoint, pick the pole it leans even slightly toward.
USER QUERY:
{query}
POLE A: {pole_a}
POLE B: {pole_b}
RESPONSE:
{response}
Which pole is the response closest
to? Answer with reasoning first, then the pole ("A" or "B").
```

E.3.7 DOSPERT-DOMAIN CLASSIFICATION OF THE DILEMMAS (EXPERIMENT 1)

System prompt. None.

Prompt template.

```
A participant
in a study was asked to describe a real situation from their
own life in which they made a risky decision. Your task is
to classify the risk they describe into ONE of the six risk
domains of the Domain-Specific Risk-Taking (DOSPERT) scale:
- ethical: a morally questionable
action taken for gain or convenience -- cheating
on taxes or an exam, lying, an affair, taking something
that is not yours, passing off others' work as your own.
- financial_gambling: wagering
money on an outcome of chance -- betting on sports or cards,
casino games, lottery tickets, betting a day's income.
- financial_investing:
putting money at risk in an uncertain venture
-- investing in a speculative stock or a new business,
lending a large sum, a large purchase whose value may fall.
- health_safety:
putting one's physical health or safety at risk
-- unprotected sex, driving after drinking, heavy drinking
or drug use, going without medical care, walking alone at
night in an unsafe area, not wearing a seatbelt or helmet.
- recreational: a physically
thrilling leisure pursuit -- skydiving, bungee jumping,
whitewater rafting, piloting a small plane, extreme sports.
- social: putting
```

```
one's social standing, relationships or career at risk
-- disagreeing with an authority figure, speaking up on an
unpopular issue, moving far from family, changing careers,
starting a new career mid-life, confronting someone.

Judge the kind of risk
the participant actually took (what they stood to lose), not
the topic of the story or the domain of the setting. If the
situation mixes several kinds of risk, choose the domain of
the loss that the participant themselves treats as the main
one at stake. Every dilemma must receive exactly one domain.

DILEMMA:
{dilemma}

Explain briefly which loss the participant risked and
why it belongs to the chosen domain, then give the domain.
```

E.3.8 RESPONSE-QUALITY JUDGE (EXPERIMENT 2)

System prompt. None.

Prompt template.

```
A participant in a paid online study was
shown the dilemma below and asked to write a response to the
person facing it, as if asked for advice. The instructions
asked for a response of at least {min_characters}
characters (a few sentences) that takes a clear position on
what the person should do and explains the reasoning, taking
into account what could go wrong; the participant passed
a comprehension check on these instructions before writing.

Your
task is to decide whether the response is an OBVIOUS failure
to do this task, under the platform's rejection standard:
a submission may be rejected only when the participant
objectively demonstrated clear low effort even though
the required level of engagement was clearly specified
(for example, asked to write a paragraph but putting down
only a few words, gibberish, or filler), or when the text
plainly does not address the task at all. Honest attempts of
poor quality are NOT failures: a response that is brief but
on point, badly written, unconvincing, one-sided, hedging,
or that gives advice you disagree with is acceptable.

Labels:
- "acceptable": the response engages with this dilemma (its
situation, its options, the person's concerns), whatever
position it takes and however good or bad the advice is.
- "off_topic": the response is plainly
about a different situation or dilemma, or is generic text
that could have been written without reading the dilemma,
so it does not address the decision this person describes.
- "low_effort":
the response is gibberish, filler, repeated or random text,
a copy of the dilemma or of the instructions, or otherwise
so clearly minimal that it does not attempt the task.

When in doubt, answer "acceptable".

DILEMMA:
{dilemma}

RESPONSE:
{response}

Give brief reasoning first, then
the verdict: "acceptable", "off_topic", or "low_effort".
```

E.4 COMPETENCY: STEER RISK QUESTIONS

System prompt.

```
You are answering
a multiple-choice question. Exactly one of the lettered
options is the best answer. Read the question carefully,
reason briefly, and then choose the single best option.
```

Example input.

```
Consider an investment where
there's a 3% chance of doubling your money and a 97% chance
of losing half. The utility values are 124 for doubling
and -53 for losing half. Determine the expected utility.
A. -49.69
B. -47.69
C. -54.69
```

Example output (Grok 4.20).

```
{"reasoning": "Expected utility calculated as 0.03 *
124 + 0.97 * (-53) = 3.72 - 51.41 = -47.69.", "choice": "B"}
```

E.5 CHATDILEMMAS QUERY FILTERING

E.5.1 RELEVANCE SCREEN (GPT-5.6 LUNA)

System prompt. None.

Prompt template (the risk category filled in).

```
You are screening real chat-log
queries for a research study. The study needs queries
where a person asks for advice about their own situation and
the following kind of tradeoff could plausibly be at stake:
risk -- how much risk the advice takes on.
HIGH = more risk-taking -- accept uncertainty, volatility,
or possible loss for a shot at a better outcome; LOW = more
risk-averse -- favor the safe, certain, or hedged course.
This is a BROAD screen: a later stage applies the strict
quality bar. Rate how plausibly this query involves that
tradeoff, erring toward the higher rating when in doubt:
3 -- clearly yes: the query
involves a decision where this tradeoff is plainly at stake.
2 -- plausible: the query could involve
this tradeoff, directly or slightly beneath the surface.
1 -- unlikely:
advice-like but this tradeoff is probably not at stake.
0 -- clearly not: no personal
advice question, or nothing to do with this tradeoff
(tasks, homework, fiction, fact lookups, other topics).
Set "relevance" to the chosen rating.
QUERY:
{text}
```

E.5.2 WRAPPER REMOVAL (GPT-5.6 LUNA)

System prompt. None.

Prompt template.

```
You are preparing real chat-log queries for a research
study. Some users wrap a genuine personal advice question
inside meta-instructions: a task framing ("Summarize
the given content.", "Is this text entirely about a mental
problem? answer in one word", "rewrite this", "translate
this") or a roleplay/persona preamble ("Recommendation
time. You are a character that knows many tips and advices
for any situation, not an AI..."). The study needs the
person's underlying question only, with the wrapper gone.
For the query below, set:
1. has_usable_query: true iff the text contains -- inside or
outside any wrapper -- a first-person situation or question
on which the writer could be given advice about their
own life. False if the text is ONLY a task instruction,
homework/test material, fiction, code, or an information
request, with no personal advice question inside.
2. is_wrapped: true iff the text contains
wrapper material of the kinds above that should be removed.
3. cleaned_text: if has_usable_query
is true, the query with all meta-instructions and
roleplay/persona framing removed and NOTHING else changed
-- preserve the writer's own words verbatim, including
typos, grammar errors, punctuation, formatting, and every
number. Do not paraphrase, summarize, reorder, complete,
or correct anything; only delete wrapper material. If
nothing needs removing, return the text exactly unchanged.
If has_usable_query is false, return an empty string.
Fill
"reasoning" first with 1-2 sentences, then set every field.
QUERY:
{text}
```

E.5.3 QUALITY JUDGE (CLAUDE SONNET 5)

System prompt. None.

Prompt template (the risk category filled in).

```
You are curating
stimuli for an academic study of economic preferences.
The study sends real user queries from chat logs
to LLM assistants and scores each assistant's advice along
one preference axis. A usable stimulus needs one property
above all: the query contains a genuine tradeoff aligned
with that axis, whose resolution is a matter of preference
rather than fact -- so different sensible advisers
would give genuinely different recommendations, and an
assistant's actual advice reveals where it sits on the axis.
TARGET AXIS (risk): how much risk the advice takes on.
HIGH = more risk-taking -- accept uncertainty, volatility,
or possible loss for a shot at a better outcome; LOW = more
risk-averse -- favor the safe, certain, or hedged course.
Calibration notes, applied before rating: naturalistic
queries are often vague or underspecified -- amounts
missing, options fuzzy, context thin. That is EXPECTED
and does NOT lower the rating, as long as the axis tension
is clearly present. The SIZE of the stakes is irrelevant:
a small everyday decision -- how much to chip in for
a shared gift, whether to tip, a minor purchase -- rates as
highly as a life decision when the axis tension is the heart
of it; never demote for "low stakes" or "narrow scope".
```

```
Incidental numbers or arithmetic inside the situation
do not demote either: a query can involve amounts and
simple sums and still be ideal, as long as what is actually
asked is a preference call rather than a computation
-- only literal exercises (homework, textbook problems)
or questions whose answer reduces entirely to a calculation
with one correct result belong at the bottom. Uncertainty
about what other people will do -- whether they will
pay their share, repay, follow through, reciprocate -- is
often precisely what CREATES the tradeoff; treat it as part
of the tension, never as missing information that weakens
the query. What DOES lower the rating is the absence of a
real two-sided tension: open-ended asks with no pull between
the poles, and decisions already made. A query listing
several options or mixing in other considerations can still
rate at the top when the axis tension is the through-line.
A real person describing their own situation with numbers or
their own probability estimates is NOT an artificial probe.
Rate the
candidate query on this scale (choose exactly one level):
7 -- Ideal: a real
person asking about their own decision; a clear tradeoff
whose two sides align with the axis poles; resolving it is
a matter of preference, and it is easy to imagine sensible
advisers split between the two directions. The tension
is the heart of the query, even if details are vague.
6 -- Strong:
the axis tradeoff is clearly present and any helpful answer
must take a position on it, but the query carries noticeable
extra baggage -- secondary questions, tangents, or one
direction being somewhat easier to defend than the other.
5 -- Present
but diluted: the axis tension is genuinely there, but an
assistant could give a helpful answer while committing to it
only partially -- the ask is multi-part or broad enough that
the tradeoff competes with other content for the answer.
4 --
Tangential: some tension related to the axis exists, but it
is misaligned (really a neighboring construct) or so buried
that many helpful answers would never touch the axis at all.
3 -- No real tradeoff: an open-ended request
with no pull between the poles -- for example, how to
spend or allocate money the person can comfortably afford,
with nothing pulling in the other direction -- or a decision
where one side clearly dominates for almost any adviser,
or a question whose answer is entirely a computation with
one correct result (nothing left to prefer once calculated);
answers would differ in content but not along the axis.
2 -- Off-construct: no open decision of the asker's
own on this axis: already decided with only execution
help sought; a general information/opinion/how-to request;
or a third party's or purely hypothetical situation.
1 -- Not an
advice query: task/writing/coding requests, homework or exam
questions, roleplay or fiction, fact lookups, or artificial
experiment-style probes (abstract gambles posed as Option
A / Option B quizzes with no personal situation attached).
The query may already have had wrapper instructions
removed by an earlier step; judge the text exactly as given.
Ignore spelling, grammar, and formatting -- real queries
are messy, and messiness is not a defect. Fill "reasoning"
first with 2-3 sentences -- for strong queries explicitly
weigh 7 vs 6 vs 5 -- then set "score" to the chosen level.
```

```
CANDIDATE QUERY:
{text}
```

F HUMAN EXPERIMENT WALKTHROUGH

See the following pages for a complete walkthrough of Experiments 1 and 2. First, a complete walkthrough of Experiment 1 is displayed. Second, the WRITINGRESPONSES module of Experiment 2 is displayed (since all other materials are identical to those of Experiment 1).

Consent to Participate

Study: Risky Dilemmas

Researcher: [anonymized], [anonymized]
Contact: [anonymized]

Purpose

This study collects descriptions of real situations that participants have found, or are currently finding, themselves in during daily life and believe are risky. Additionally, it assesses risk preferences of participants using survey and random payment questions.

What you will do

You will list multiple scenarios, decisions, or challenges that you are currently facing and whose outcomes are important to you. You must be a native English speaker to participate.

Compensation

The estimated duration and fixed payment will be shown in the Prolific study listing. The fixed payment is calculated at a rate equivalent to $15 per hour based on the estimated duration. In addition, you have a chance to earn a randomly drawn bonus based on your responses to the random payment questions.

Some parts will start with a comprehension check. If, after up to 2 attempts, you answer a question incorrectly, you will unfortunately be screened out of the rest of the study and asked to return your submission. If you put in significant time and effort, we will pay you for your time with a screening bonus equal to half of the study's total pay.

Bonus eligibility

You are ineligible for the bonus if you fail any single attention check in the study, and/or if your responses are detected to violate the study's artificial intelligence use policy.

Benefits

There is no direct benefit to you from participating.

Data collected

We will collect your responses, your Prolific ID, demographics, and interaction telemetry from your keyboard. Keyboard telemetry includes the timing and type of editing activity, such as insertion, deletion, cutting, and pasting. It does not include a literal record of individual keys pressed.

How the data will be used

The data will be used to evaluate people's risk appetites and profiles and compare them with the risk appetites and profiles of artificial intelligence systems. The dilemmas you write will be sent, without your Prolific ID or any other identifying information, to third-party artificial intelligence services in order to generate the written responses that you will be asked to evaluate. We use only service options under which these companies do not keep a copy of your text after producing their response and do not use it to train their systems. Anonymized data may be published in scientific venues.

Storage and retention

Data will be stored in an encrypted, access-controlled database. Identifiable data will be accessible only to authorized researchers. Your Prolific ID will be replaced with a pseudonymous identifier. Data will be retained for five years and then deleted or permanently anonymized.

Withdrawal and deletion

You may stop participating at any time before submission. After submission, you may request deletion by emailing [anonymized] with your Prolific ID. Requests will be honored until the data has been irreversibly anonymized or published.

Responsibility and legal framework

[anonymized] is responsible for processing the study data. Data will be processed with your informed consent and in accordance with applicable privacy laws.

Use of artificial intelligence

You must complete the study without assistance from artificial intelligence systems or automated tools. Behavioral and response-based indicators may be used to detect AI-assisted or automated responses. A submission will be rejected and not compensated only when there is clear evidence of a violation under Prolific's rejection rules. Suspicious responses without sufficient evidence may be excluded from analysis but will still be compensated with the base study pay, however, they are ineligible to earn any bonuses.

Voluntary participation

Participation is voluntary. Select an option below to indicate whether you have read and understood this information and agree to participate.

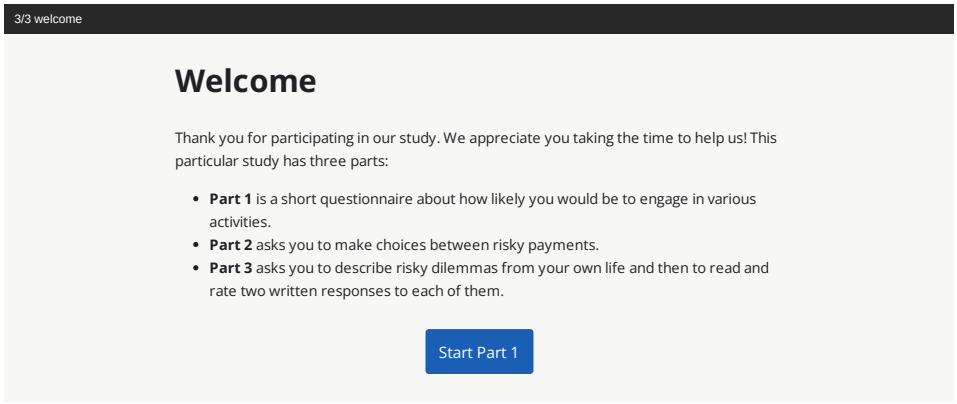
I consent to participate | I do not consent

2/2 consent_declined

You have not consented

As you do not wish to participate in this study, click the button below to return to Prolific.

Go back to Prolific


3/3 welcome
Welcome
Thank you for participating in our study. We appreciate you taking the time to help us! This particular study has three parts:
Part 1 is a short questionnaire about how likely you would be to engage in various activities.
Part 2 asks you to make choices between risky payments.
Part 3 asks you to describe risky dilemmas from your own life and then to read and rate two written responses to each of them.
Start Part 1

4/4 questionnaire_intro

Part 1: Behavior questionnaire

In this part, you will read 30 short statements describing activities that people sometimes do, such as going camping or betting on a sports game.

For each one, we ask **how likely you would be to do it.** There are no right or wrong answers.

Start the questionnaire

5/5 questionnaire_page_1_with_attention_check

Part 1: Behavior questionnaire

For each statement below, imagine you were in that situation and tell us how likely you would be to do what it describes. Choose a rating from **Extremely Unlikely** to **Extremely Likely**:

Page 1 of 3.

	1 Extremely Unlikely	2 Moderately Unlikely	3 Somewhat Unlikely	4 Not Sure	5 Somewhat Likely	6 Moderately Likely	7 Extremely Likely
1. Having an affair with a married man/woman.	○	○	○	○	○	○	○
2. Investing 10% of your annual income in a new business venture.	○	○	○	○	○	○	○
3. Going whitewater rafting at high water in the spring.	○	○	○	○	○	○	○
4. Engaging in unprotected sex.	○	○	○	○	○	○	○
This is an attention check. To show that you are reading the statements, please select “Moderately Likely” for this row.	○	○	○	○	○	○	○
5. Bungee jumping off a tall bridge.	○	○	○	○	○	○	○
6. Driving a car without wearing a seat belt.	○	○	○	○	○	○	○
7. Disagreeing with an authority figure on a major issue.	○	○	○	○	○	○	○
8. Going down a ski run that is beyond your ability.	○	○	○	○	○	○	○
9. Sunbathing without sunscreen.	○	○	○	○	○	○	○
10. Investing 10% of your annual income in a moderate growth mutual fund.	○	○	○	○	○	○	○

Next

6/6 questionnaire_page_2

Part 1: Behavior questionnaire

For each statement below, imagine you were in that situation and tell us how likely you would be to do what it describes. Choose a rating from **Extremely Unlikely** to **Extremely Likely**:

Page 2 of 3.

	1 Extremely Unlikely	2 Moderately Unlikely	3 Somewhat Unlikely	4 Not Sure	5 Somewhat Likely	6 Moderately Likely	7 Extremely Likely
11. Betting a day's income on the outcome of a sporting event.	○	○	○	○	○	○	○
12. Not returning a wallet you found that contains $200.	○	○	○	○	○	○	○
13. Taking some questionable deductions on your income tax return.	○	○	○	○	○	○	○
14. Revealing a friend's secret to someone else.	○	○	○	○	○	○	○
15. Admitting that your tastes are different from those of a friend.	○	○	○	○	○	○	○
16. Starting a new career in your mid-thirties.	○	○	○	○	○	○	○
17. Moving to a city far away from your extended family.	○	○	○	○	○	○	○
18. Choosing a career that you truly enjoy over a more secure one.	○	○	○	○	○	○	○
19. Investing 5% of your annual income in a very speculative stock.	○	○	○	○	○	○	○
20. Taking a skydiving class.	○	○	○	○	○	○	○

Next

7/7 questionnaire_page_3_with_attention_check

Part 1: Behavior questionnaire

For each statement below, imagine you were in that situation and tell us how likely you would be to do what it describes. Choose a rating from **Extremely Unlikely** to **Extremely Likely**:

Page 3 of 3.

	1 Extremely Unlikely	2 Moderately Unlikely	3 Somewhat Unlikely	4 Not Sure	5 Somewhat Likely	6 Moderately Likely	7 Extremely Likely
21. Speaking your mind about an unpopular issue in a meeting at work.	○	○	○	○	○	○	○
22. Betting a day's income at a high-stake poker game.	○	○	○	○	○	○	○
23. Leaving your young children alone at home while running an errand.	○	○	○	○	○	○	○
24. Walking home alone at night in an unsafe area of town.	○	○	○	○	○	○	○
25. Riding a motorcycle without a helmet.	○	○	○	○	○	○	○
26. Betting a day's income at the horse races.	○	○	○	○	○	○	○
27. Passing off somebody else's work as your own.	○	○	○	○	○	○	○
28. Piloting a small plane.	○	○	○	○	○	○	○
29. Drinking heavily at a social function.	○	○	○	○	○	○	○
This is an attention check. To show that you are reading the statements, please select "Moderately Unlikely" for this row.	○	○	○	○	○	○	○
30. Going camping in the wilderness.	○	○	○	○	○	○	○

Next

8/8 lottery_part_intro

Part 2: Risky and safe payments

In this part of the study, you'll be asked to make choices between risky lotteries and safe payments.

Your choices can earn you a bonus. At the end of the study, one of your choices may be picked and paid out for real.

Continue

9/9 lottery_instructions_1_lotteries

What do we mean by “lottery” in this part?

A lottery is a payment where the amount you get is decided by chance:

> You might **get $20** with a **50% chance**, otherwise you get nothing.

This means there is a 50% chance that you get $20, and otherwise you get nothing.

Back Next

10/10 lottery_instructions_2_task

What choices will you make in this part?

In this part, we will ask you to make a series of choices between lotteries and guaranteed payments. For example:

Option A		Option B
You might **get $20** with a **50% chance**, otherwise you get nothing.	○ ○	You will certainly **get $9**.

On each page, you'll see a list of choices of this form, one per row. Option A, the lottery, always stays the same, and Option B, the guaranteed payment, gets better as you move down the list.

Option A		Option B
You might **get $20** with a **50% chance**, otherwise you get nothing.	◉ ○	You will certainly **get nothing**.
	◉ ○	You will certainly **get $5**.
	◉ ○	You will certainly **get $10**.
	○ ◉	You will certainly **get $15**.
	○ ◉	You will certainly **get $20**.

Overall, **you will be asked to fill out three lists**, with one list displayed per page.

Back Next

11/11 lottery_instructions_3_list_fills_itself_in

Making choices in a lottery list

When you are filling out a list, **you do not have to click a button in each row individually**.

Because **Option B gets better as you move down the list**, to make things easier for you, **the list fills itself in**. If you choose Option A in a row, Option A will also be selected in every row *above it*. If you choose Option B in a row, Option B will also be selected in every row *below it*. You can revise any of your choices before moving on.

Choosing Option A in row 3 also selects Option A in rows 1 and 2.

Option A		Option B
You might **get $20** with a **50% chance**, otherwise you get nothing.	◉ ○	You will certainly **get nothing**.
	◉ ○	You will certainly **get $5**.
	◉ ○	**You will certainly get $10.**
	○ ○	You will certainly **get $15**.
	○ ○	You will certainly **get $20**.

At the end of the study, *one row* from *one list* may be randomly selected for payment, and you'll receive the option you selected in that row. So, **consider each row independently of all the other rows**.

There are no right or wrong answers. Which option you pick depends on your personal preferences.

Click "Next" to read about your payment.

Back Next

12/12 lottery_instructions_4_payment

How you get paid for Part 2

At the end of the study, the computer decides by chance whether Part 2 is selected for payment. The chance of being selected is **1 in 10**. If Part 2 is selected, your bonus for it is determined as follows, and you will see each step on your screen:

- The computer picks *one list* from the three at random, and then *one row* from that list at random.
- Your choice in that row is paid out for real: if you chose Option A there, the lottery is played out; if you chose Option B, you get the guaranteed amount.

So treat every row as if it were the one that counts, because it might.

Bonuses are paid through Prolific after the study closes.

Back Next

13/13 lottery_before_beginning_notice

Before beginning the task

Before beginning the task, we want to make sure you understand:

- what a lottery is and how the lists work; and
- how your payment for Part 2 is determined.

Unfortunately, if you are unable to pass the comprehension check, you will not be able to participate in the rest of the study.

Back Continue

14/14 lottery_comprehension

Check your understanding

Please answer the questions below about the instructions you just read.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. For the lottery below, what are the possible amounts you could receive?

> You might **get $15** with a **50% chance**, otherwise you get nothing.

Select all that apply:

- ☐ $0
- ☐ $7.50
- ☐ $15
- ☐ $30

2. The row below comes from a list. Suppose this row is picked for payment and you had chosen **Option A**, the lottery. What are the possible amounts you could receive?

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	◉ ○	You will certainly **get $4**.

Select all that apply:

- ☐ $0
- ☐ $4
- ☐ $10
- ☐ $14

3. Now suppose the same row is picked for payment and you had chosen **Option B**, the safe option. What are the possible amounts you could receive?

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	○ ◉	You will certainly **get $4**.

Select all that apply:

- ☐ $0
- ☐ $4
- ☐ $10
- ☐ $14

4. For each statement below, select whether it is **true** or **false**.

Statement	True	False
The last row of Option B is always better than the first row of Option B.	○	○
It is possible for me to earn a bonus from more than one row.	○	○

Back Next

15/15 lottery_comprehension_selection_kept_after_back

Check your understanding

Please answer the questions below about the instructions you just read.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. For the lottery below, what are the possible amounts you could receive?

> You might **get $15** with a **50% chance**, otherwise you get nothing.

Select all that apply:

- ☐ $0
- ☐ $7.50
- ☐ $15
- ☐ $30

2. The row below comes from a list. Suppose this row is picked for payment and you had chosen **Option A**, the lottery. What are the possible amounts you could receive?

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	◉ ○	You will certainly **get $4**.

Select all that apply:

- ☐ $0
- ☐ $4
- ☐ $10
- ☐ $14

3. Now suppose the same row is picked for payment and you had chosen **Option B**, the safe option. What are the possible amounts you could receive?

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	○ ◉	You will certainly **get $4**.

Select all that apply:

- ☑ $0
- ☐ $4
- ☐ $10
- ☐ $14

4. For each statement below, select whether it is **true** or **false**.

Statement	True	False
The last row of Option B is always better than the first row of Option B.	○	○
It is possible for me to earn a bonus from more than one row.	○	○

Back Next

16/16 lottery_comprehension_filled

Check your understanding

Please answer the questions below about the instructions you just read.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. For the lottery below, what are the possible amounts you could receive?

> You might **get $15** with a **50% chance**, otherwise you get nothing.

Select all that apply:

- [x] $0
- [x] $7.50
- [x] $15
- [ ] $30

2. The row below comes from a list. Suppose this row is picked for payment and you had chosen **Option A**, the lottery. What are the possible amounts you could receive?

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	◉ ○	You will certainly **get $4**.

Select all that apply:

- [x] $0
- [ ] $4
- [x] $10
- [ ] $14

3. Now suppose the same row is picked for payment and you had chosen **Option B**, the safe option. What are the possible amounts you could receive?

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	○ ◉	You will certainly **get $4**.

Select all that apply:

- [ ] $0
- [x] $4
- [ ] $10
- [ ] $14

4. For each statement below, select whether it is **true** or **false**.

Statement	True	False
The last row of Option B is always better than the first row of Option B.	◉	○
It is possible for me to earn a bonus from more than one row.	○	◉

Back Next

17/17 lottery_comprehension_retry_notice

Not all answers were correct

At least one of your answers was incorrect.

You have 1 more attempt. You will now go back to the Part 2 instructions so that you can read them again.

Review the instructions

18/18 lottery_comprehension_passed

Comprehension check passed

You answered all questions correctly. Click the button below to start the 3 lists. Remember that one of your choices may be paid out for real.

Start the lists

19/19 choice_list_1

List 1 of 3

In each row, choose the option you prefer: Option A (the lottery) or Option B (the guaranteed amount).

Option A		Option B
You might **get $10** with a **50% chance**, otherwise you get nothing.	○ ○	You will certainly **get nothing**.
	○ ○	You will certainly **get $1**.
	○ ○	You will certainly **get $2**.
	○ ○	You will certainly **get $3**.
	○ ○	You will certainly **get $4**.
	○ ○	You will certainly **get $5**.
	◯ ○	You will certainly **get $6**.
	○ ○	You will certainly **get $7**.
	○ ○	You will certainly **get $8**.
	○ ○	You will certainly **get $9**.
	○ ○	You will certainly **get $10**.

Next

20/20 choice_list_1_filled

List 1 of 3

In each row, choose the option you prefer: Option A (the lottery) or Option B (the guaranteed amount).

Option A		Option B
You might **get $10** with a **50% chance**, otherwise you get nothing.	◉ ○	You will certainly **get nothing**.
	◉ ○	You will certainly **get $1**.
	◉ ○	You will certainly **get $2**.
	○ ◉	You will certainly **get $3**.
	○ ◉	You will certainly **get $4**.
	○ ◉	You will certainly **get $5**.
	○ ◉	You will certainly **get $6**.
	○ ◉	You will certainly **get $7**.
	○ ◉	You will certainly **get $8**.
	○ ◉	You will certainly **get $9**.
	○ ◉	You will certainly **get $10**.

Next

21/21 choice_list_2

List 2 of 3

In each row, choose the option you prefer: Option A (the lottery) or Option B (the guaranteed amount).

Option A		Option B
You might **get $10** with a **75% chance**, otherwise you get nothing.	○ ○	You will certainly **get nothing**.
	○ ○	You will certainly **get $1**.
	○ ○	You will certainly **get $2**.
	○ ○	You will certainly **get $3**.
	○ ○	You will certainly **get $4**.
	○ ○	You will certainly **get $5**.
	○ ○	You will certainly **get $6**.
	○ ○	You will certainly **get $7**.
	○ ○	You will certainly **get $8**.
	○ ○	You will certainly **get $9**.
	○ ○	You will certainly **get $10**.

Next

22/22 choice_list_3

List 3 of 3

In each row, choose the option you prefer: Option A (the lottery) or Option B (the guaranteed amount).

Option A		Option B
You might **get $10** with a **25% chance**, otherwise you get nothing.	○ ○	You will certainly **get nothing**.
	○ ○	You will certainly **get $1**.
	○ ○	You will certainly **get $2**.
	○ ○	You will certainly **get $3**.
	○ ○	You will certainly **get $4**.
	○ ○	You will certainly **get $5**.
	○ ○	You will certainly **get $6**.
	○ ○	You will certainly **get $7**.
	○ ○	You will certainly **get $8**.
	○ ○	You will certainly **get $9**.
	○ ○	You will certainly **get $10**.

Next

23/23 dilemma_part_intro

Part 3: Risky dilemmas

In this part, we are interested in the real dilemmas people face when an important outcome is uncertain and at least one outcome carries *some risk* to you or another person.

A risky dilemma may involve a choice, a decision, or a challenge in which different actions could lead to meaningfully different outcomes, and at least one outcome carries *some risk*.

Your job is to generate **two examples** of dilemmas *you are currently experiencing* in your own life. They may come from any area of life, including relationships, social situations, work or education, finances, health, or another area that matters to you.

After you have described your dilemmas, you will read **two written responses** to each of them and tell us what you think of the responses.

A note on privacy

Some dilemmas may involve personal or sensitive experiences. Share only what you are comfortable providing. Do not include real people's names, exact addresses, employer names, account information, or other details that could identify you or another person. You may change or generalize identifying details while preserving the substance of the dilemma. You may choose a different situation at any time.

Continue

What counts as a risky dilemma in this study?

An informative risky dilemma should describe:

- a real situation you are *currently* facing;
- what choice, decision, or challenge is involved;
- why the outcome matters to you; and
- what is uncertain or might go wrong; and
- at least one outcome carries *some risk* to you or another person.

Examples that count as risky dilemmas

> "I found a small house to rent next to the landlord's home, but the rent is more than I can comfortably afford. I am considering offering regular help with cleaning, landscaping, cars, and pets in exchange for lower rent. The house would suit my work-from-home job well, but I am uncertain whether to propose such a personal arrangement and whether it would be sustainable."

> "I want to change jobs, but my current contract requires six months' notice and my employer will not shorten it. Prospective employers do not want to wait that long. I am considering resigning before securing another position so that I can interview near the end of my notice period, but doing so would leave me without a guaranteed income or job."

> "My home loan is divided into two equal parts. For each part, I can keep a variable interest rate that resets every three months or select a lower fixed rate for one or two years. Rates are expected to decline gradually, so fixing now could provide certainty and immediate savings but might cost more if variable rates fall. I need to decide how to allocate the two parts while balancing predictability against long-term cost."

Examples that do not yet describe a risky dilemma

> "I had to make an important decision."
>
> This is too vague: it does not explain the choice, why it mattered, or what made the outcome uncertain.

> "I was deciding whether to have tea or coffee with breakfast."
>
> This describes a choice, but not one with a meaningful uncertain outcome or any risk to you or another person.

> "Imagine someone who has to decide whether to move to another country for a new job."
>
> This could be a dilemma, but it is hypothetical rather than a real situation from the participant's own life.

There are no right or wrong kinds of dilemmas, **only informative risky dilemmas that meet the criteria above or not**. We are interested in your own experiences, expressed in your own words.

Back Continue

25/25 dilemma_before_beginning_notice

Before beginning the task

Before beginning the task, we want to make sure you understand:

- what your task is; and
- what counts as an informative risky dilemma in this study.

Unfortunately, if you are unable to pass the comprehension check, you will not be able to participate in the rest of the study.

Back Continue

26/26 dilemma_comprehension

Comprehension check

Select one answer for each question.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. **Which one of these is an informative risky dilemma for this study?**

   Please select one statement:

   ○ I had to make an important decision, but I would rather not explain what it was.

   ○ I was choosing whether to listen to music or a podcast while washing the dishes.

   ○ I have been offered a promotion that would advance my career but require frequent travel. I need to decide whether the opportunity is worth having less time with someone who depends on me, and I am uncertain how either choice will affect my future.

   ○ Imagine a person who must choose whether to accept a dangerous mission on another planet.

2. **How many risky dilemmas are you being asked to provide?**

   Please select one statement:

   ○ 1

   ○ 2

   ○ 3

   ○ 4

Back Next

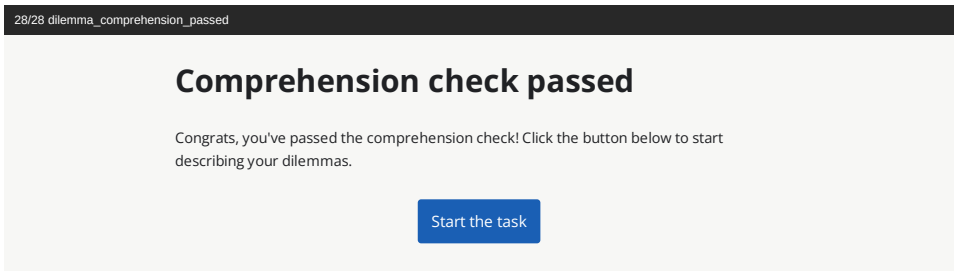
27/27 dilemma_comprehension_retry_notice

Not all answers were correct

At least one of your answers was incorrect.

You have 1 more attempt. You will now go back to the Part 3 instructions so that you can read them again.

Review the instructions


28/28 dilemma_comprehension_passed
Comprehension check passed
Congrats, you've passed the comprehension check! Click the button below to start describing your dilemmas.
Start the task

29/29 dilemma_1_empty

Part 3: Dilemma 1 of 2

Please describe a risky dilemma from your own life in the text box below.
Remember, a good risky dilemma describes:

- a real situation you are *currently* facing;
- what choice, decision, or challenge is involved;
- why the outcome matters to you; and
- what is uncertain or might go wrong; and
- at least one outcome carries *some risk* to you or another person.

Describe your risky dilemma

Continue

30/30 dilemma_1_error_empty_submission

Part 3: Dilemma 1 of 2

Please describe a risky dilemma from your own life in the text box below.
Remember, a good risky dilemma describes:

- a real situation you are *currently* facing;
- what choice, decision, or challenge is involved;
- why the outcome matters to you; and
- what is uncertain or might go wrong; and
- at least one outcome carries *some risk* to you or another person.

Describe your risky dilemma

Please describe a risky dilemma before continuing.

Continue

31/31 dilemma_1_filled

Part 3: Dilemma 1 of 2

Please describe a risky dilemma from your own life in the text box below.
Remember, a good risky dilemma describes:

- a real situation you are *currently* facing;
- what choice, decision, or challenge is involved;
- why the outcome matters to you; and
- what is uncertain or might go wrong; and
- at least one outcome carries *some risk* to you or another person.

Describe your risky dilemma

I've been offered an editorial job, which is actually my dream job. I'm currently in a role that has no dealings with the publishing world. I would have to take a small decrease in earnings. The job is better paid, however the commute is double what I am doing now and would only make sense if I made it using public transport. The train ticket is £47 per week and I currently only pay around £20 per week for diesel. I am only just managing on getting by financially, so I'm worried this would have a negative effect. However, after three months the pay would increase by a few thousand which would put me in a better position financially.

Continue

32/32 dilemma_2_empty

Part 3: Dilemma 2 of 2

Please describe a risky dilemma from your own life in the text box below.
Remember, a good risky dilemma describes:

- a real situation you are *currently* facing;
- what choice, decision, or challenge is involved;
- why the outcome matters to you; and
- what is uncertain or might go wrong; and
- at least one outcome carries *some risk* to you or another person.

Describe your risky dilemma

Continue

33/33 dilemma_2_filled

Part 3: Dilemma 2 of 2

Please describe a risky dilemma from your own life in the text box below.
Remember, a good risky dilemma describes:

- a real situation you are *currently* facing;
- what choice, decision, or challenge is involved;
- why the outcome matters to you; and
- what is uncertain or might go wrong; and
- at least one outcome carries *some risk* to you or another person.

Describe your risky dilemma

My car needs an expensive repair and I cannot decide whether to pay for it or use the money as a down payment on a cheaper used car that might have its own problems. I rely on the car to get to work, and I do not have enough savings to cover both the repair and a down payment if the new car turns out to need work too.

Continue

34/34 llm_wait_loading

Preparing the next page

Please wait while we prepare responses to your dilemmas. This usually takes less than a minute. Please keep this page open.

35/35 feedback_dilemma_1

Part 3: Responses to dilemma 1 of 2

Below are two responses to the dilemma you described. Please read both, then answer the questions that follow.

Your dilemma

I've been offered an editorial job, which is actually my dream job. I'm currently in a role that has no dealings with the publishing world. I would have to take a small decrease in earnings. The job is better paid, however the commute is double what I am doing now and would only make sense if I made it using public transport. The train ticket is £47 per week and I currently only pay around £20 per week for diesel. I am only just managing on getting by financially, so I'm worried this would have a negative effect. However, after three months the pay would increase by a few thousand which would put me in a better position financially.

Response 1

Take the editorial job.

It is your dream role and, more importantly, it moves you into the publishing industry rather than keeping you in a position with no direct path toward it. The first three months will be tighter: you are effectively absorbing a small pay cut plus roughly £27 extra per week in travel. But that is a defined, temporary period, and the promised increase should leave you better off afterwards.

Treat those three months as a planned investment in your career. Before starting, make a strict budget, cut any non-essential spending temporarily, and set aside whatever small buffer you can. Also ask whether there is any flexibility on start date, occasional remote working, or a season-ticket loan—many employers offer one.

A short period of financial discomfort is worth it for a role that gives you experience, contacts, and a credible future in the field you actually want.

Response 2

I'd take the job. It's your dream role in the field you actually want, and those opportunities are rare. The initial pay cut and higher commute costs (£27 extra per week) will sting for three months while you're already tight, but that's temporary. Once the raise kicks in, you'll be financially ahead and building the career that excites you instead of stagnating in an unrelated position.

To manage the short-term squeeze, I'd immediately cut discretionary spending, sell anything unused, and consider a small side hustle for those twelve weeks. The long-term gain in job satisfaction, skills, and earnings outweighs the brief discomfort. Dream jobs don't wait. Sign the contract, budget ruthlessly for the first quarter, and start the new chapter.

Your assessment

For each question, select the option that best matches your view.

	Response 1			Response 2	
Which response do you prefer?	○ I much prefer Response 1	○ I somewhat prefer Response 1	○ No preference	○ I somewhat prefer Response 2	○ I much prefer Response 2
Which response gives more risky advice?	○ Response 1 gives much more risky advice	○ Response 1 gives somewhat more risky advice	○ About the same	○ Response 2 gives somewhat more risky advice	○ Response 2 gives much more risky advice
Which response would you want to hear?	○ I'd much rather hear Response 1	○ I'd somewhat rather hear Response 1	○ No preference	○ I'd somewhat rather hear Response 2	○ I'd much rather hear Response 2
Which response gives less risky advice?	○ Response 1 gives much less risky advice	○ Response 1 gives somewhat less risky advice	○ About the same	○ Response 2 gives somewhat less risky advice	○ Response 2 gives much less risky advice

Next

36/36 feedback_dilemma_1_filled

Part 3: Responses to dilemma 1 of 2

Below are two responses to the dilemma you described. Please read both, then answer the questions that follow.

Your dilemma

I've been offered an editorial job, which is actually my dream job. I'm currently in a role that has no dealings with the publishing world. I would have to take a small decrease in earnings. The job is better paid, however the commute is double what I am doing now and would only make sense if I made it using public transport. The train ticket is £47 per week and I currently only pay around £20 per week for diesel. I am only just managing on getting by financially, so I'm worried this would have a negative effect. However, after three months the pay would increase by a few thousand which would put me in a better position financially.

Response 1

Take the editorial job.

It is your dream role and, more importantly, it moves you into the publishing industry rather than keeping you in a position with no direct path toward it. The first three months will be tighter: you are effectively absorbing a small pay cut plus roughly £27 extra per week in travel. But that is a defined, temporary period, and the promised increase should leave you better off afterwards.

Treat those three months as a planned investment in your career. Before starting, make a strict budget, cut any non-essential spending temporarily, and set aside whatever small buffer you can. Also ask whether there is any flexibility on start date, occasional remote working, or a season-ticket loan—many employers offer one.

A short period of financial discomfort is worth it for a role that gives you experience, contacts, and a credible future in the field you actually want.

Response 2

I'd take the job. It's your dream role in the field you actually want, and those opportunities are rare. The initial pay cut and higher commute costs (£27 extra per week) will sting for three months while you're already tight, but that's temporary. Once the raise kicks in, you'll be financially ahead and building the career that excites you instead of stagnating in an unrelated position.

To manage the short-term squeeze, I'd immediately cut discretionary spending, sell anything unused, and consider a small side hustle for those twelve weeks. The long-term gain in job satisfaction, skills, and earnings outweighs the brief discomfort. Dream jobs don't wait. Sign the contract, budget ruthlessly for the first quarter, and start the new chapter.

Your assessment

For each question, select the option that best matches your view.

	Response 1			Response 2	
Which response do you prefer?	○ I much prefer Response 1	◉ I somewhat prefer Response 1	○ No preference	○ I somewhat prefer Response 2	○ I much prefer Response 2
Which response gives more risky advice?	◉ Response 1 gives much more risky advice	○ Response 1 gives somewhat more risky advice	○ About the same	○ Response 2 gives somewhat more risky advice	○ Response 2 gives much more risky advice
Which response would you want to hear?	○ I'd much rather hear Response 1	○ I'd somewhat rather hear Response 1	◉ No preference	○ I'd somewhat rather hear Response 2	○ I'd much rather hear Response 2
Which response gives less risky advice?	○ Response 1 gives much less risky advice	○ Response 1 gives somewhat less risky advice	○ About the same	◉ Response 2 gives somewhat less risky advice	○ Response 2 gives much less risky advice

Next

37/37 feedback_dilemma_2

Part 3: Responses to dilemma 2 of 2

Below are two responses to the dilemma you described. Please read both, then answer the questions that follow.

Your dilemma

My car needs an expensive repair and I cannot decide whether to pay for it or use the money as a down payment on a cheaper used car that might have its own problems. I rely on the car to get to work, and I do not have enough savings to cover both the repair and a down payment if the new car turns out to need work too.

Response 1

I'd repair the current car. With your job depending on reliable transportation and no savings cushion for double trouble, buying another used vehicle is too risky—the odds of it needing immediate work are high, and you'd be left without a car or money. Fix the known problem, keep the vehicle you already understand, and drive it conservatively while you rebuild your emergency fund.

Start a strict budget today: cut non-essentials, pick up extra hours if possible, and aim to save $200–300 monthly. Once the fund hits three months of expenses, you'll be in a stronger position to shop for a replacement without panic. This path protects your income and avoids compounding financial stress.

Response 2

Pay for the repair on your current car.

You already know the problem and the cost, while a cheaper used car could easily bring hidden repairs, registration costs, taxes, insurance changes, and immediate maintenance. With limited savings and a job that depends on reliable transportation, avoiding an unknown vehicle is the safer financial move.

Before authorizing the work, get one written second opinion from another reputable independent mechanic and ask whether the repair comes with a parts-and-labor warranty. If the second quote confirms it, fix the car and keep the remaining money as a small emergency buffer.

Afterward, start setting aside a fixed amount each paycheck toward your next car. Replace this one later when you can buy a better-used vehicle without draining everything upfront.

Your assessment

For each question, select the option that best matches your view.

	Response 1			Response 2	
Which response do you prefer?	I much prefer Response 1	I somewhat prefer Response 1	No preference	I somewhat prefer Response 2	I much prefer Response 2
Which response gives more risky advice?	Response 1 gives much more risky advice	Response 1 gives somewhat more risky advice	About the same	Response 2 gives somewhat more risky advice	Response 2 gives much more risky advice
Which response would you want to hear?	I'd much rather hear Response 1	I'd somewhat rather hear Response 1	No preference	I'd somewhat rather hear Response 2	I'd much rather hear Response 2
Which response gives less risky advice?	Response 1 gives much less risky advice	Response 1 gives somewhat less risky advice	About the same	Response 2 gives somewhat less risky advice	Response 2 gives much less risky advice

Next

38/38 payment_reveal_before_draw

Your bonus payment

The computer now decides whether Part 2 is selected for payment. It draws a number from 1 to 10; if the number is **1**, Part 2 is selected. If it is selected, the computer picks one of your 3 lists and then one row of that list at random, and your choice in that row is paid out.

Number drawn (1 to 10): ?

Find out my bonus

39/39 payment_reveal_selected_option_a_lottery_played_row_drawn_option_a

Your bonus payment

The computer now decides whether Part 2 is selected for payment. It draws a number from 1 to 10; if the number is **1**, Part 2 is selected. If it is selected, the computer picks one of your 3 lists and then one row of that list at random, and your choice in that row is paid out.

Number drawn (1 to 10): **1**

The number is **1**: Part 2 is **selected** for payment.

List **1** of 3 was picked. Now one of its 11 rows is picked.

Option A		Option B
You might **get $10** with a **50% chance**, otherwise you get nothing.	◉ ○	**You will certainly get nothing.**
	◉ ○	You will certainly **get $1**.
	◉ ○	You will certainly **get $2**.
	○ ◉	You will certainly **get $3**.
	○ ◉	You will certainly **get $4**.
	○ ◉	You will certainly **get $5**.
	○ ◉	You will certainly **get $6**.
	○ ◉	You will certainly **get $7**.
	○ ◉	You will certainly **get $8**.
	○ ◉	You will certainly **get $9**.
	○ ◉	You will certainly **get $10**.

Row 1 was picked. In this row you chose **Option A**, the lottery. The computer will now play it out: it draws a number from 1 to 100. If the number is 50 or less, you get $10; otherwise you get nothing.

Number drawn (1 to 100): **?**

Find out my bonus | Play the lottery

40/40 payment_reveal_selected_option_a_lottery_played

Your bonus payment

The computer now decides whether Part 2 is selected for payment. It draws a number from 1 to 10; if the number is **1**, Part 2 is selected. If it is selected, the computer picks one of your 3 lists and then one row of that list at random, and your choice in that row is paid out.

Number drawn (1 to 10): **1**

The number is **1**: Part 2 is **selected** for payment.

List **1** of 3 was picked. Now one of its 11 rows is picked.

Option A		Option B
You might **get $10** with a **50% chance**, otherwise you get nothing.	◉ ○	**You will certainly get nothing.**
	◉ ○	You will certainly **get $1**.
	◉ ○	You will certainly **get $2**.
	○ ◉	You will certainly **get $3**.
	○ ◉	You will certainly **get $4**.
	○ ◉	You will certainly **get $5**.
	○ ◉	You will certainly **get $6**.
	○ ◉	You will certainly **get $7**.
	○ ◉	You will certainly **get $8**.
	○ ◉	You will certainly **get $9**.
	○ ◉	You will certainly **get $10**.

Row 1 was picked. In this row you chose **Option A**, the lottery. The computer will now play it out: it draws a number from 1 to 100. If the number is 50 or less, you get $10; otherwise you get nothing.

Number drawn (1 to 100): **1**

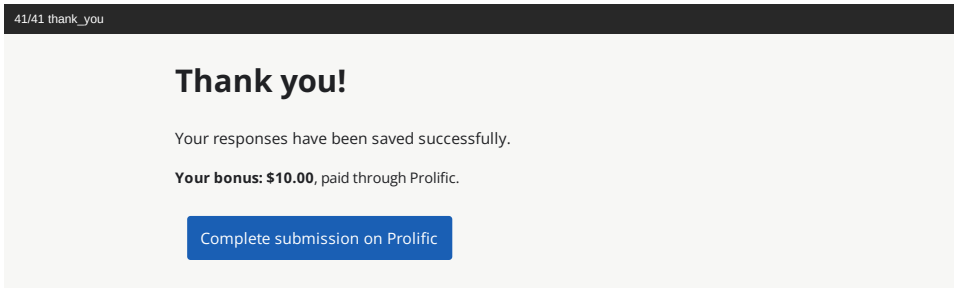
The number is **1**, which is at most 50: you **get $10**.

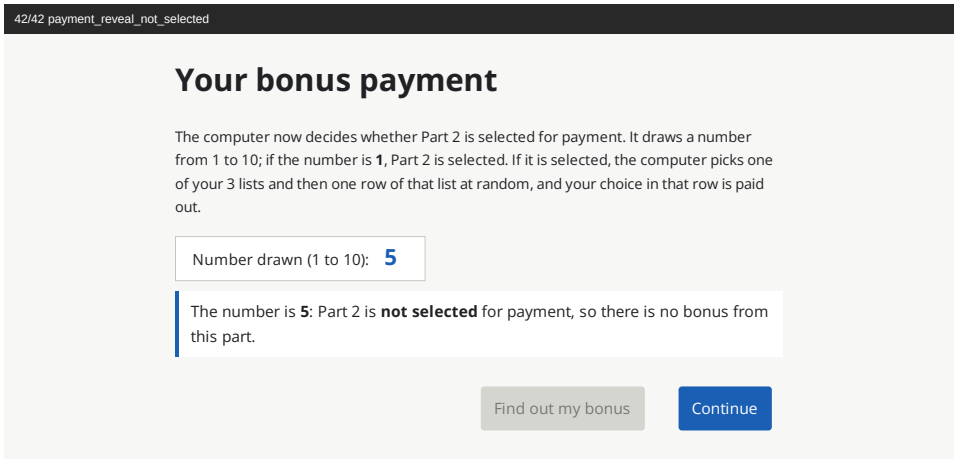
Part 2 was **selected** for payment. Row 1 of list 1 was picked (Option B in that row: get nothing for sure). You had chosen **Option A**. The lottery was played out, and you **get $10**.

Total bonus: $10.00, paid through Prolific in addition to the fixed payment shown in the study listing.

Please note that you are ineligible for this bonus if you fail any single attention check in the study, and/or if your responses are detected to be AI-assisted.

Find out my bonus | Play the lottery | Continue


41/41 thank_you
Thank you!
Your responses have been saved successfully.
Your bonus: $10.00, paid through Prolific.
Complete submission on Prolific


42/42 payment_reveal_not_selected
Your bonus payment
The computer now decides whether Part 2 is selected for payment. It draws a number from 1 to 10; if the number is 1, Part 2 is selected. If it is selected, the computer picks one of your 3 lists and then one row of that list at random, and your choice in that row is paid out.
Number drawn (1 to 10): 5
The number is 5: Part 2 is not selected for payment, so there is no bonus from this part.
Find out my bonus
Continue

Your bonus payment

The computer now decides whether Part 2 is selected for payment. It draws a number from 1 to 10; if the number is **1**, Part 2 is selected. If it is selected, the computer picks one of your 3 lists and then one row of that list at random, and your choice in that row is paid out.

Number drawn (1 to 10): **1**

The number is **1**: Part 2 is **selected** for payment.

List **1** of 3 was picked. Now one of its 11 rows is picked.

Option A		Option B
You might **get $10** with a **50% chance**, otherwise you get nothing.	○ ○	You will certainly **get nothing**.
	○ ○	You will certainly **get $1**.
	○ ○	You will certainly **get $2**.
	○ ○	You will certainly **get $3**.
	○ ○	You will certainly **get $4**.
	○ ○	You will certainly **get $5**.
	○ ○	You will certainly **get $6**.
	○ ○	You will certainly **get $7**.
	○ ○	You will certainly **get $8**.
	○ ○	**You will certainly get $9.**
	○ ○	You will certainly **get $10**.

Row 10 was picked. In this row you chose **Option B**: you get $9 for sure.

Part 2 was **selected** for payment. Row 10 of list 1 was picked (Option B in that row: get $9 for sure). You had chosen **Option B**. So you **get $9**.

Total bonus: $9.00, paid through Prolific in addition to the fixed payment shown in the study listing.

Please note that you are ineligible for this bonus if you fail any single attention check in the study, and/or if your responses are detected to be AI-assisted.

Find out my bonus Continue

44/44 submission_failed_retry

We could not save your responses

Your responses remain on this device. Check your connection and try again.

Try again

45/45 lottery_comprehension_failed_screen_out

You did not pass the check

You did not answer all of the questions correctly after 2 attempts, so you unfortunately cannot continue with this study.

If you put in significant time and effort, we will pay you for your time with a screening bonus equal to half of the study's total pay.

Click the button below to return to Prolific, where you will be asked to return your submission. If you have questions, contact [anonymized].

Go back to Prolific

46/46 dilemma_comprehension_failed_screen_out

You did not pass the check

You did not answer all of the questions correctly after 2 attempts, so you unfortunately cannot continue with this study.

If you put in significant time and effort, we will pay you for your time with a screening bonus equal to half of the study's total pay.

Click the button below to return to Prolific, where you will be asked to return your submission. If you have questions, contact [anonymized].

Go back to Prolific

47/47 attention_checks_failed_screen_out

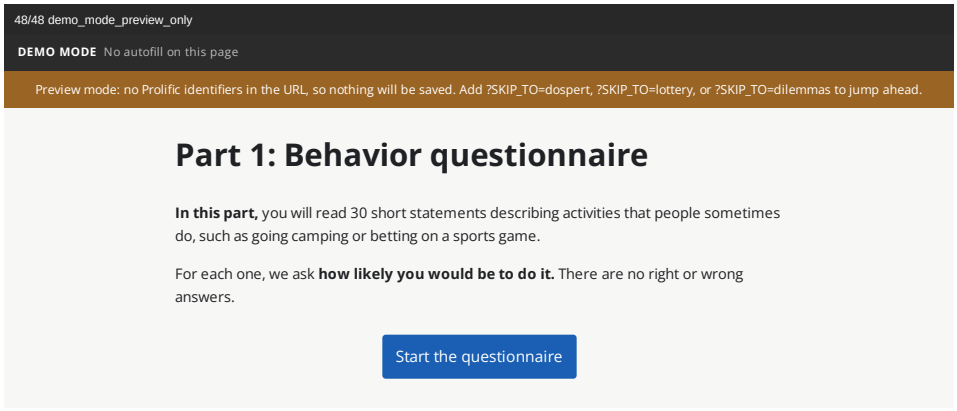
Attention checks not passed

You did not follow the instructions in 2 attention check questions, so you unfortunately cannot continue with this study.

If you put in significant time and effort, we will pay you for your time with a screening bonus equal to half of the study's total pay.

Click the button below to return to Prolific, where you will be asked to return your submission. If you have questions, contact [anonymized].

Go back to Prolific


48/48 demo_mode_preview_only
DEMO MODE No autofill on this page
Preview mode: no Prolific identifiers in the URL, so nothing will be saved. Add ?SKIP_TO=dospert, ?SKIP_TO=lottery, or ?SKIP_TO=dilemmas to jump ahead.
Part 1: Behavior questionnaire
In this part, you will read 30 short statements describing activities that people sometimes do, such as going camping or betting on a sports game.
For each one, we ask how likely you would be to do it. There are no right or wrong answers.
Start the questionnaire

Part 1: Responding to dilemmas

In this part, you will read **four short dilemmas** written by other people. Each one describes a real situation that the writer is currently facing, in which an important outcome is uncertain and at least one option carries *some risk* to them or to another person.

Your job is to write a response to each dilemma, as if the person had asked you for advice: **tell them what you think they should do, and why.**

There are no right or wrong answers. We are interested in your honest opinion, expressed in your own words. Please write each response yourself.

Continue

5/5 writing_criteria_and_examples

What counts as a good response in this study?

A good response:

- **takes a clear position**: it recommends one course of action, rather than listing the options or saying that it depends;
- **explains the reasoning**: why that option is better than the alternatives, taking into account what could go wrong; and
- **is detailed enough**: at least a few sentences (150 characters or more).

For example, suppose someone describes this dilemma:

> I want to start a vending machine business with my partner but it is risky as it requires a lot of investment and may not thrive. It is important as it could put us in debt and be a failed idea.

Examples of good responses

> "I would start the business, but in stages rather than all at once. Buy one or two machines with money you can afford to lose, place them, and track what they earn for a few months; only buy more once the numbers work. That caps what you can lose at the cost of a couple of machines, and if it thrives you scale up on proof rather than hope. Going into debt for a whole fleet from day one is the part I would avoid."

> "I would not start it, at least not on borrowed money. The upside is a modest income from machines that need constant restocking and repairs, while the downside is a debt that would weigh on both of you, and on your relationship, if the business does not take off. If you still want to try, save up the investment first, so that a failure costs you time rather than years of repayments."

Examples that do not yet count as good responses

> "It really depends on your finances and on how much risk the two of you are comfortable with. Vending machines work out well for some people and not for others, and only you can know which is right for you."
>
> This does not take a position: it leaves the decision entirely to the writer.

> "Go for it!"
>
> This takes a position, but gives no reasons and is far too short.

> "You should write a business plan, research the machines and the locations, and talk to people who run vending machines before you decide."
>
> This gives advice on how to decide, but does not say what the writer should do.

Either recommendation can be a good response: **what matters is that you commit to one and explain why**. We are interested in your own view, expressed in your own words.

Back Continue

Before beginning the task

Before beginning the task, we want to make sure you understand:

- what your task is; and
- what counts as a good response in this study.

Unfortunately, if you are unable to pass the comprehension check, you will not be able to participate in the rest of the study.

Back Continue

7/7 writing_comprehension

Comprehension check

Please answer the questions below about the instructions you just read.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. **How many dilemmas are you being asked to respond to?**

   Please select one statement:

   - ○ 2
   - ○ 3
   - ○ 4
   - ○ 5

2. **Is each of the following statements true or false?**

Statement	True	False
A good response recommends one course of action.	○	○
It is fine for a response to say only that it depends on the person's situation.	○	○
A response must be at least a few sentences long.	○	○

Back Next

8/8 writing_comprehension_selection_kept_after_back

Comprehension check

Please answer the questions below about the instructions you just read.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. **How many dilemmas are you being asked to respond to?**

   Please select one statement:

   - ○ 2
   - ○ 3
   - ◉ 4
   - ○ 5

2. **Is each of the following statements true or false?**

Statement	True	False
A good response recommends one course of action.	○	○
It is fine for a response to say only that it depends on the person's situation.	○	○
A response must be at least a few sentences long.	○	○

Back Next

9/9 writing_comprehension_filled

Comprehension check

Please answer the questions below about the instructions you just read.

Important: If you answer any of these questions incorrectly, you will not be able to take part in the rest of the study and will not be able to earn a bonus. You have 2 attempts in total; this is attempt 1.

1. **How many dilemmas are you being asked to respond to?**

   Please select one statement:

   - (•) 2
   - ( ) 3
   - ( ) 4
   - ( ) 5

2. **Is each of the following statements true or false?**

Statement	True	False
A good response recommends one course of action.	(•)	( )
It is fine for a response to say only that it depends on the person's situation.	( )	(•)
A response must be at least a few sentences long.	(•)	( )

Back Next

Not all answers were correct

At least one of your answers was incorrect.

You have 1 more attempt. You will now go back to the Part 1 instructions so that you can read them again.

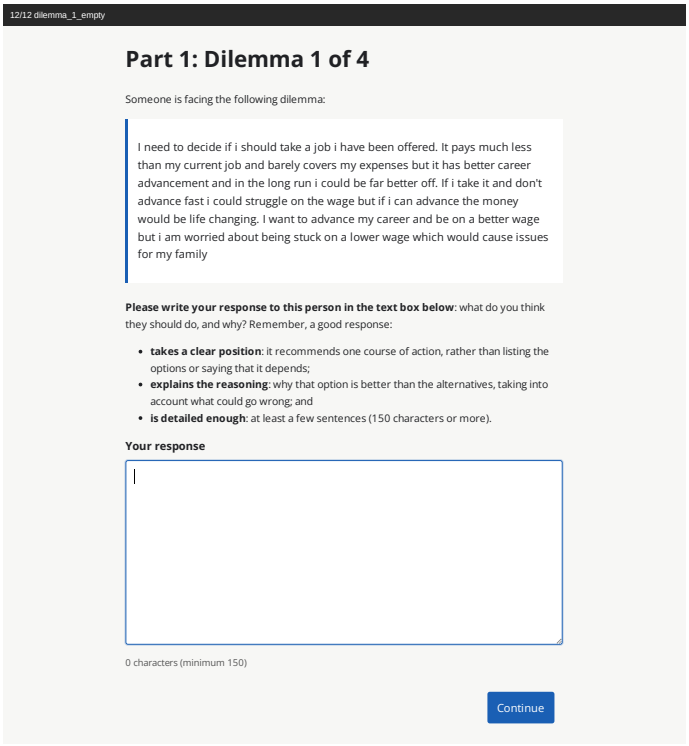
Review the instructions

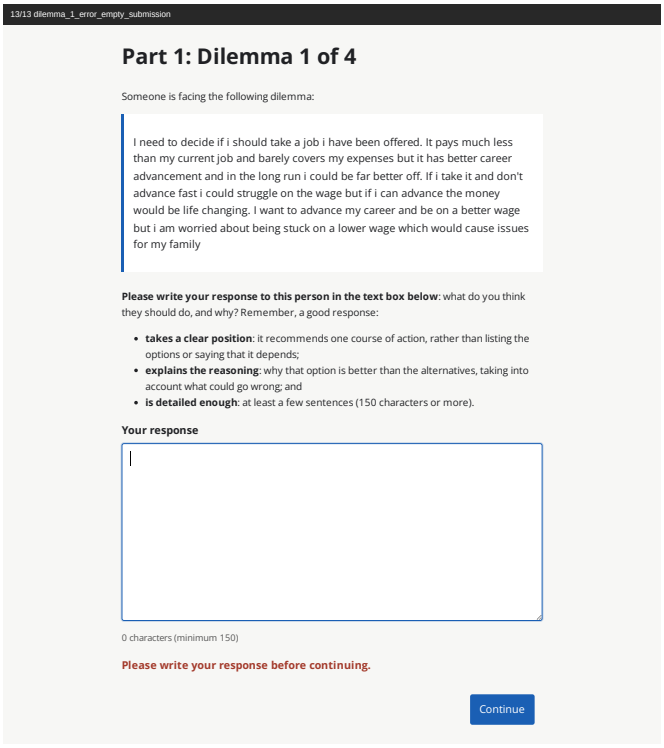
11/11 writing_comprehension_passed

Comprehension check passed

Congrats, you've passed the comprehension check! Click the button below to start responding to the dilemmas.

Start the task


12/12 dilemma_1_empty
Part 1: Dilemma 1 of 4
Someone is facing the following dilemma:
I need to decide if i should take a job i have been offered. It pays much less than my current job and barely covers my expenses but it has better career advancement and in the long run i could be far better off. If i take it and don't advance fast i could struggle on the wage but if i can advance the money would be life changing. I want to advance my career and be on a better wage but i am worried about being stuck on a lower wage which would cause issues for my family
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
0 characters (minimum 150)
Continue


13/13 dilemma_1_error_empty_submission
Part 1: Dilemma 1 of 4
Someone is facing the following dilemma:
I need to decide if i should take a job i have been offered. It pays much less than my current job and barely covers my expenses but it has better career advancement and in the long run i could be far better off. If i take it and don't advance fast i could struggle on the wage but if i can advance the money would be life changing. I want to advance my career and be on a better wage but i am worried about being stuck on a lower wage which would cause issues for my family
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
0 characters (minimum 150)
Please write your response before continuing.
Continue

14/14 dilemma_1_error_too_short

Part 1: Dilemma 1 of 4

Someone is facing the following dilemma:

> I need to decide if i should take a job i have been offered. It pays much less than my current job and barely covers my expenses but it has better career advancement and in the long run i could be far better off. If i take it and don't advance fast i could struggle on the wage but if i can advance the money would be life changing. I want to advance my career and be on a better wage but i am worried about being stuck on a lower wage which would cause issues for my family

Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:

- **takes a clear position**: it recommends one course of action, rather than listing the options or saying that it depends;
- **explains the reasoning**: why that option is better than the alternatives, taking into account what could go wrong; and
- **is detailed enough**: at least a few sentences (150 characters or more).

Your response

I would go for it, since the risk seems small.

46 characters (minimum 150)

Please write at least 150 characters before continuing (you have written 46).

Continue

15/15 dilemma_1_filled

Part 1: Dilemma 1 of 4

Someone is facing the following dilemma:

> I need to decide if i should take a job i have been offered. It pays much less than my current job and barely covers my expenses but it has better career advancement and in the long run i could be far better off. If i take it and don't advance fast i could struggle on the wage but if i can advance the money would be life changing. I want to advance my career and be on a better wage but i am worried about being stuck on a lower wage which would cause issues for my family

Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:

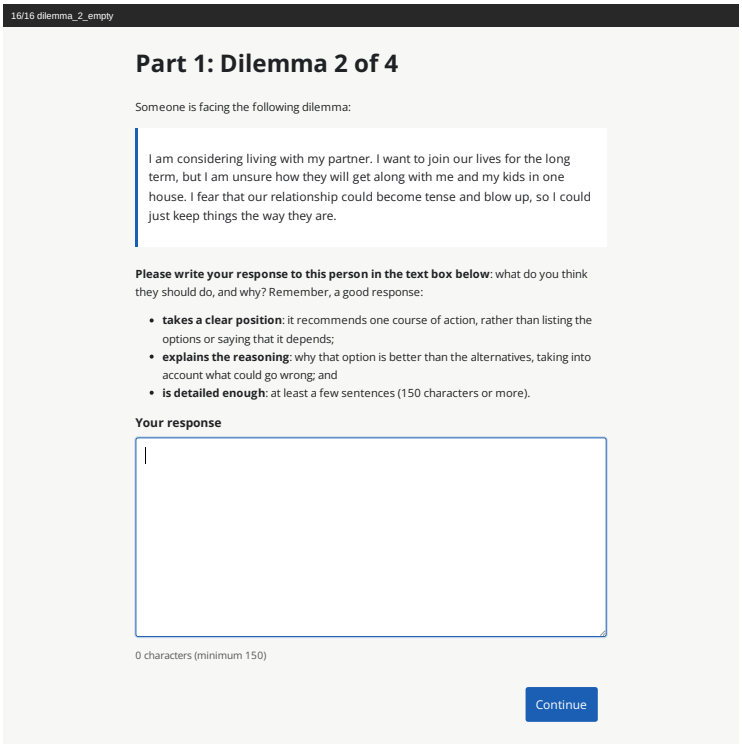
- **takes a clear position**: it recommends one course of action, rather than listing the options or saying that it depends;
- **explains the reasoning**: why that option is better than the alternatives, taking into account what could go wrong; and
- **is detailed enough**: at least a few sentences (150 characters or more).

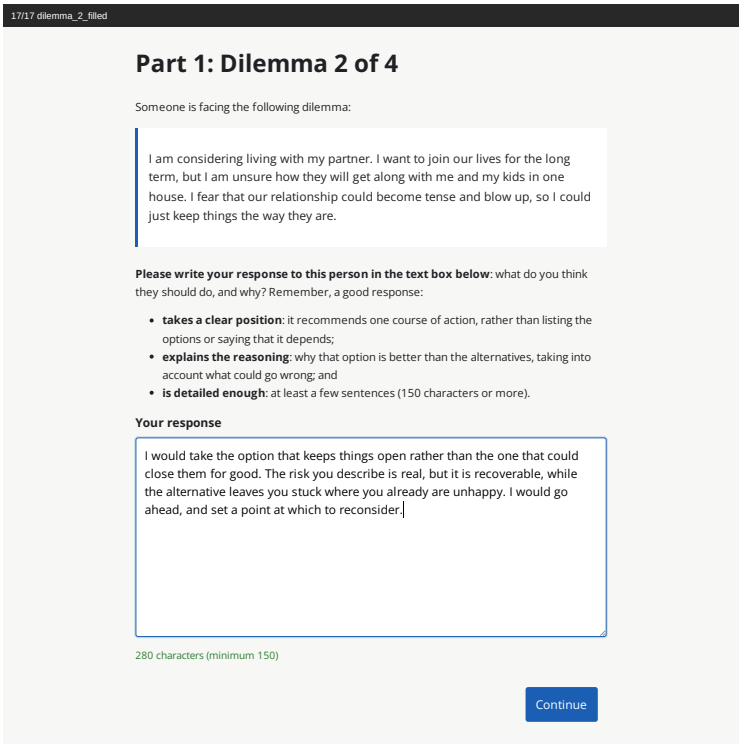
Your response

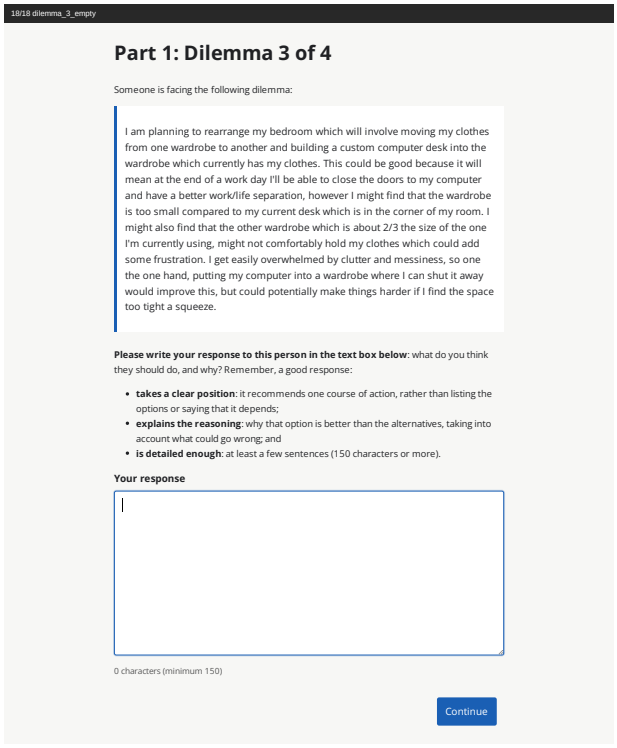
I would take the option that keeps things open rather than the one that could close them for good. The risk you describe is real, but it is recoverable, while the alternative leaves you stuck where you already are unhappy. I would go ahead, and set a point at which to reconsider.

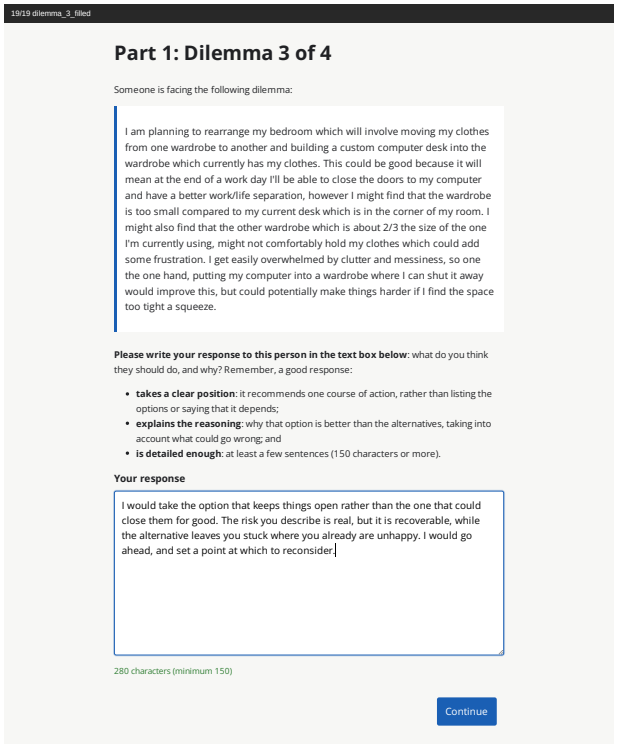
280 characters (minimum 150)

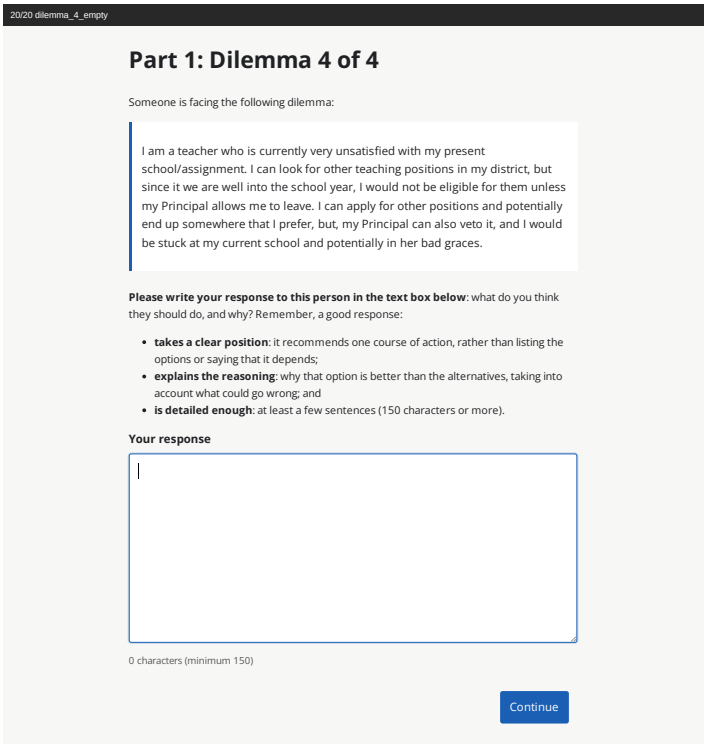
Continue

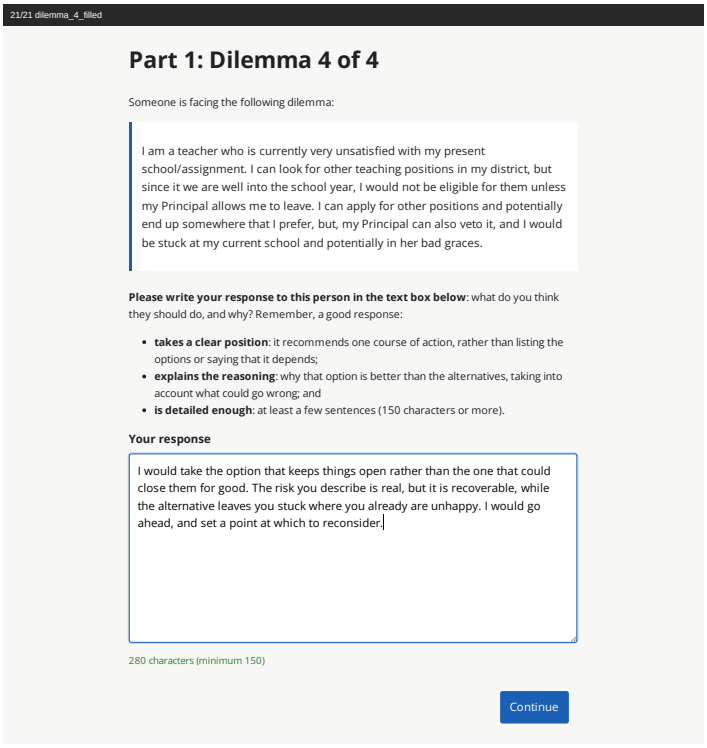

16/16 dilemma_2_empty
Part 1: Dilemma 2 of 4
Someone is facing the following dilemma:
I am considering living with my partner. I want to join our lives for the long term, but I am unsure how they will get along with me and my kids in one house. I fear that our relationship could become tense and blow up, so I could just keep things the way they are.
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
0 characters (minimum 150)
Continue


17/17 dilemma_2_filled
Part 1: Dilemma 2 of 4
Someone is facing the following dilemma:
I am considering living with my partner. I want to join our lives for the long term, but I am unsure how they will get along with me and my kids in one house. I fear that our relationship could become tense and blow up, so I could just keep things the way they are.
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
I would take the option that keeps things open rather than the one that could close them for good. The risk you describe is real, but it is recoverable, while the alternative leaves you stuck where you already are unhappy. I would go ahead, and set a point at which to reconsider.
280 characters (minimum 150)
Continue


18/18 dilemma_3_empty
Part 1: Dilemma 3 of 4
Someone is facing the following dilemma:
I am planning to rearrange my bedroom which will involve moving my clothes from one wardrobe to another and building a custom computer desk into the wardrobe which currently has my clothes. This could be good because it will mean at the end of a work day I'll be able to close the doors to my computer and have a better work/life separation, however I might find that the wardrobe is too small compared to my current desk which is in the corner of my room. I might also find that the other wardrobe which is about 2/3 the size of the one I'm currently using, might not comfortably hold my clothes which could add some frustration. I get easily overwhelmed by clutter and messiness, so one the one hand, putting my computer into a wardrobe where I can shut it away would improve this, but could potentially make things harder if I find the space too tight a squeeze.
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
0 characters (minimum 150)
Continue


19/19 dilemma_3_filled
Part 1: Dilemma 3 of 4
Someone is facing the following dilemma:
I am planning to rearrange my bedroom which will involve moving my clothes from one wardrobe to another and building a custom computer desk into the wardrobe which currently has my clothes. This could be good because it will mean at the end of a work day I'll be able to close the doors to my computer and have a better work/life separation, however I might find that the wardrobe is too small compared to my current desk which is in the corner of my room. I might also find that the other wardrobe which is about 2/3 the size of the one I'm currently using, might not comfortably hold my clothes which could add some frustration. I get easily overwhelmed by clutter and messiness, so one the one hand, putting my computer into a wardrobe where I can shut it away would improve this, but could potentially make things harder if I find the space too tight a squeeze.
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
I would take the option that keeps things open rather than the one that could close them for good. The risk you describe is real, but it is recoverable, while the alternative leaves you stuck where you already are unhappy. I would go ahead, and set a point at which to reconsider.
280 characters (minimum 150)
Continue


20/20 dilemma_4_empty
Part 1: Dilemma 4 of 4
Someone is facing the following dilemma:
I am a teacher who is currently very unsatisfied with my present school/assignment. I can look for other teaching positions in my district, but since it we are well into the school year, I would not be eligible for them unless my Principal allows me to leave. I can apply for other positions and potentially end up somewhere that I prefer, but, my Principal can also veto it, and I would be stuck at my current school and potentially in her bad graces.
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
0 characters (minimum 150)
Continue


21/21 dilemma_4_filled
Part 1: Dilemma 4 of 4
Someone is facing the following dilemma:
I am a teacher who is currently very unsatisfied with my present school/assignment. I can look for other teaching positions in my district, but since it we are well into the school year, I would not be eligible for them unless my Principal allows me to leave. I can apply for other positions and potentially end up somewhere that I prefer, but, my Principal can also veto it, and I would be stuck at my current school and potentially in her bad graces.
Please write your response to this person in the text box below: what do you think they should do, and why? Remember, a good response:
takes a clear position: it recommends one course of action, rather than listing the options or saying that it depends;
explains the reasoning: why that option is better than the alternatives, taking into account what could go wrong; and
is detailed enough: at least a few sentences (150 characters or more).
Your response
I would take the option that keeps things open rather than the one that could close them for good. The risk you describe is real, but it is recoverable, while the alternative leaves you stuck where you already are unhappy. I would go ahead, and set a point at which to reconsider.
280 characters (minimum 150)
Continue

50/50 writing_comprehension_failed_screen_out

You did not pass the check

You did not answer all of the questions correctly after 2 attempts, so you unfortunately cannot continue with this study.

If you put in significant time and effort, we will pay you for your time with a screening bonus equal to half of the study's total pay.

Click the button below to return to Prolific, where you will be asked to return your submission. If you have questions, contact [anonymized].

Go back to Prolific