Designing behavioral interventions with explainable artificial intelligence

Anonymous CogSci submission

Abstract

Identifying how language drives human behavior is crucial, both practically to influence people and theoretically for understanding psychology, culture, and myriad behaviors predicated on verbal or written communication. Here, we present a method based on explainable artificial intelligence (AI) that identifies what aspects of the language people read are associated with their reactions. As an example, we use this method to redesign choice tasks to modulate people's risk taking across five novel datasets containing 38,911 responses to 255 distinct decision tasks. Overall, we both i) design interventions to influence behavior at scale with greater efficiency than methods such as AB-testing and ii) extract previously unknown drivers of behavior as a basis for developing new theories across academic disciplines.

Keywords: risky decision making; explainable AI; behavioral interventions; sentence embeddings

Introduction

The field of psychology offers a clear example of the importance of language, focusing largely on studying proximal, behavioral responses to stimuli that are frequently language based. In a typical research design, scholars manipulate individual elements of the language used in the stimuli and observe changes in the corresponding responses to uncover the psychological mechanisms underpinning behavior. The chosen manipulations are designed using individual researchers' expertise and psychological theory. Though language may not necessarily be the underlying construct of interest—which could be anything from judgment to personality—it is a primary means by which such constructs are manipulated.

The paradigm of research described above has been enormously generative, but it has historically faced two key limitations: i) the requirement that researchers choose particular manipulations to test can lead to gaps of awareness, and ii) the use of factorial designs—where one or more dimensions are varied orthogonally—imposes a resource constraint on how many dimensions can be explored at the same time. With the advent of new developments in natural language processing and artificial intelligence (AI) explainability techniques, such limitations no longer hold true. With these modern methods, language can be represented as numbers, these numbers can be used to predict behavior, and then these predictive relationships can be interpreted in terms of their original form of language to offer theoretical insights. We describe the two main technical advancements that facilitate this possibility next.

Artificial neural networks can represent the language of a text and its subtle semantic variations as a set of numbers in high-dimensional vector space called an "embedding" (Mikolov et al., 2013). Intuitively, one can think about embeddings as positioning words in a vector space where the different dimensions capture different aspects of meaning and the relative distances between words capture how similar their meanings are. More recent advancements that use "transformers" in their neural network architectures can capture contextualized meaning, enabling richer understanding of the meaning of phrases and sentences than earlier word-focused approaches (Vaswani et al., 2017; Naveed et al., 2024).

Yet embeddings alone do not facilitate understanding how language impacts behavior—indeed their individual dimensions are generally uninterpretable (Boselli et al., 2024; Şenel et al., 2018)—so something further is required. We show that by combining these embeddings with modern techniques from explainable AI—which focus on the question of why predictive AI models make the predictions that they do (Molnar, 2025; Linardatos et al., 2021)—it is possible to extract robust psychological insights regarding the link between language and human behavior.

In the present research, we develop a general methodology to uncover the contextual relationships between language and behavior and use these relationships to design targeted behavioral interventions. To do so, we first generate variation in the language of textual stimuli while maintaining the same informational content (e.g., outcomes, probabilities) and document people's responses to the different versions of each stimulus. We then embed these textual stimuli using large language models (LLMs), which are highly effective at producing contextualized embeddings that represent the meaning of texts (Tao et al., 2024) and predict people's responses using the text of the stimuli they saw. We use AI explainability techniques on these predictive models to unpack the relationships between the language of these texts and participants' responses. Specifically, we apply a method called "integrated gradients" (Sundararajan et al., 2017) where we simulate iteratively changing the meaning of individual words relative to a semantically neutral baseline and see how this changes the model's predictions for people's behavior. This enables us to infer which words and phrases are associated with particular types of responses. Finally, we generate new stimuli based on the insights from this method, and test whether they influence behavior as predicted in new out-of-sample tests.

This general methodology can be applied to study any relationship between language and behavior, but we demonstrate its utility in decision making under risk, which focuses on people's responses to decision contexts where potential outcomes and their associated probabilities are known (Molins et al., 2024). Such problems offer an ideal setting to study our approach as a high impact,

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

interdisciplinary topic where people make choices based on both objective and subjective features of the context and the importance of language has been highlighted.

Our work differs in important ways from prior research, which has focused on two primary lines of investigation. Scholars typically either i) use embeddings to investigate how risk perceptions are represented in language, exposing how people evaluate and perceive different naturalistic sources of risk (Bhatia, 2019), or ii) design prompts for LLMs to perform tasks such as constructing personalized messages based on personality traits to influence behavior (Matz et al., 2024) or generating reasons for and against a risky behavior that can be used to understand someone's choice to engage in that behavior (Bhatia, 2024). Our research instead focuses on how seemingly arbitrary variations in the text of a stimulus impact behavior, identifying which specific words and phrases have such effects. In doing so, we can detect relationships that may not be cleanly interpretable based on extant knowledge—similar to AlphaGo's infamous "Move 37", which did not initially make sense to Go experts.

The data we collected also facilitate further lines of investigation. Researchers have previously highlighted that stimulus-level variability in effects may reflect confounding factors undermining the hypothesized behavioral effects (Simonsohn et al., 2025). In our designs, we intentionally vary language independently of the types of features (e.g., domain) that are typically focused on. Will framing effects be robust to such changes? Identifying the extent to which framing effects emerge across many different rewritten versions of a problem can both test their robustness while also providing informative evidence regarding the extent to which they are likely to be practically meaningful in applied contexts where many factors vary simultaneously. To study this, we introduce the idea of "semantic robustness"—the extent to which a manipulation (e.g., gain-loss framing) induces a separation in outcomes between two sets of stimuli that vary in the language they use, rather than a single focal pair.

We next describe our methods for each of our three studies, before turning to reporting our results and discussion.

Methods

Study 1

We recruited 199 Prolific participants (M_{age}=41.9, 48.2% men, 49.3% women, 2.5% non-binary) to rephrase the decision-making problems. A further 97 participants were also recruited via Prolific Academic (M_{age}=42.8, 46.4% male, 49.5% female, 4.1% non-binary) to generate additional rewrites of the risk-inhibiting version of the economic risky-choice framing problem, where the initial data collection did not obtain 11 valid rewrites. We subsequently recruited 1,775 Prolific participants to answer the rewritten versions. Decision problems were presented to participants using Javascript, which led to a non-zero but low error rate. We excluded participants from the analysis who stated that there was an error in the display of the decision problems, as per our pre-registration: https://tinyurl.com/4tesydkb. Further, one participant provided data twice, leading us to exclude their second submission. In total, our final sample consisted of 1,740 participants (M_{age}=43.8, 48.7% men, 49.8% women, 1.5% other/prefer not to answer).

We chose three different kinds of framing manipulations that have been the focus of extant literature (see Levin et al., 1998). For each of these three types of framing problems, we chose one problem in the domain of health, and one in the domain of economic decision-making as important and general contexts to study. For each of the six problems, there is one frame that increases risk taking, relative to a second frame that inhibits risk taking. We refer to problem-domain combinations as "seeds" and the particular risk-inducing/inhibiting version of each problem-domain combination as a "seed frame". Thus, in total, we had 12 seed frames, which we asked Prolific participants and ChatGPT-4 to rewrite. We refer to rewrites of a seed frame as isomorphs.

The first and second author independently coded each participant-generated isomorph as either being a valid rewrite or not. Rewrites were classified as valid if the key parts of the information, meaning the potential outcomes and probabilities associated with each choice, as well as the overall context of the problem remained the same. Additionally, rewrites were only classified as valid if the framing remained consistent. Isomorphs were only kept if both authors independently coded the rewrite as valid. Overall, we collected 11 participant-based and three GPT-based isomorphs of each original text, for a total of 15 versions of each seed frame when including the originals, constituting 180 unique problems in total.

When collecting risky behavior data for these rewritten versions, participants first answered an attention check question and provided informed consent. Each participant then answered one risk inducing and one risk inhibiting decision frame for each of the six seed problems, which always entailed making binary choices between two courses of action. They answered six decision frames in the first block, which were either all risk inducing or risk inhibiting frames of the six seed problems, and vice versa for the second block. The ordering of the blocks, as well as the ordering of the six problems within a block was randomly assigned at the individual level. For each risk inducing and inhibiting frame of each seed problem, the observed isomorph was also randomly assigned at the individual level. After applying our preregistered exclusion criteria, our dataset consisted of 20,880 choices.

Following past research on within-subject framing designs (LeBoeuf & Shafir, 2003), participants answered questions unrelated to the framing contexts between the first and second block of problems, to minimize the likelihood that participants noticed the normative equivalence between the risk-inducing and risk-inhibiting frames. Specifically, participants were asked to answer a short measure of their own overconfidence using the Overconfidence Test (OCT, Lawson et al., 2024), as well as the Cognitive Reflection Test (CRT, Frederick, 2005) between the first and second block.

The ordering of the OCT and CRT blocks, as well as the ordering of the three items within each block was randomly assigned. After providing answers to all of the risky choice problems, participants answered the Domain-Specific Risk-Taking (DOSPERT) scale as a measure of their risk-taking attitudes (averaging the investment and gambling subscales, see Weber et al., 2002). Finally, participants provided demographic information including gender, age, level of education, and ethnicity, and were then debriefed.

Study 2

We recruited 1,000 participants via Prolific Academic to collect responses to the rewritten decision-making problems. After removing duplicate responses and participants with display errors as pre-registered we were left with 971 participants (M_{age}=42.0, 49.1% men, 49.2% women, 1.7% other/prefer not to answer; pre-registration here: https://tinyurl.com/2ruffxpy).

We harness Llama 3.0's ('Llama' henceforth; Grattafiori et al., 2024) learned representations by keeping its parameters fixed and training a series of ℓ2-penalized regressions (one for each seed) to predict isomorphs' observed Bernoulli probability distributions (obtained by averaging individual responses) from Llama's embeddings of isomorphs' texts. To interpret the effects of language on human choice, we use Integrated Gradients ('IG' hereafter, Sundararajan et al., 2017) to estimate how changes to the isomorph will affect the model's predicted proportion. This procedure obtains a score that approximates the contribution of each element to the obtained output, where larger positive scores imply a greater contribution toward the predicted value (i.e., greater risk-taking proportion) and vice versa. We define our baseline to be the embedding of Llama's end-of-message token as it is a semantically neutral element, and apply this computation to each isomorph's input token embeddings using a trapezoidal Riemann sum to approximate the integral with 500 steps (Kokhlikyan et al., 2020).

After attributing each token an importance score, we map the tokens back to words and obtain importance scores at the word level. Similarly, we obtain phrase importances by defining phrases as tokens appearing between consecutive punctuation elements (commas, periods, question marks, exclamation marks, colons, and semicolons), and mapping tokens to phrases. We then compute each word or phrase's final score by averaging its IG scores across the respective seed frame's isomorphs in which it appeared. We opted to source these attributions for both words and phrases since the valence of words can be altered by the particular lexical and topical context in which they appear (Polanyi & Zaenen, 2006). We similarly focus on the seed-frame level as opposed to the seed level to mitigate concerns that the attribution merely captures framing effects and to produce meaningful attributions (e.g., "saved" could be attributed a biased score since it is used positively in gain frames of risky-choice framing and negatively ("not saved") in loss frames. These opposing use cases are likely to cancel out and lead to less meaningful attributions when computed across the seed.)

We then used the attributions to craft new risk-inducing and risk-inhibiting versions of each seed frame. Specifically, we ordered the attributions from most associated with risky choice to least associated with risky choice, and constructed new versions of the seed-frame isomorph, including a risk-inducing variant (using the top ranked words/phrases to maximize their summed scores) and a risk-inhibiting variant (using the bottom-ranked words/phrases to minimize their summed scores). It was necessary to do this while also using common sense to edit for readability, and we further checked that the risk-inducing rewrite had a predicted risk propensity meaningfully exceeding that of the risk-inhibiting one before proceeding. In turn, we obtain for each seed frame four new variants: word-based risk inducing, word-based risk inhibiting, phrase-based risk inducing, phrase-based risk inhibiting.

When collecting risk preferences for these attribution-based rewrites, the experimental procedure was near identical to Study 1. The only change was that rather than answering one of 15 versions of each problem, participants were randomly assigned to a 2 (word-based vs. phrase-based) x 2 (risk-inhibiting vs. risk-inducing) mixed design. Participants viewed either phrase-based or word-based rewrites in a between-subjects assignment, and for each of the 12 rewrites that participants answered, they were randomly assigned at the individual level to viewing either a risk-inhibiting or risk-inducing frame. The rest of the procedure and randomization was identical. The final dataset consisted of 11,652 unique risky decisions made by participants.

Study 3

We developed a real-life decision task for the choice of an optional insurance for a vacation by consulting actual insurance offers. The parameters of the task (duration of the vacation, cost of the vacation, as well as cost of insurance) mirrored real data observed on Expedia, a popular online travel booking platform. The task diverges from the framing problems of Study 1 and 2, as there are no objective probabilities associated with each outcome. To source variations in the language of the task we prompted ChatGPT-4o in chat to rephrase the insurance's policy section 20 times, generating a lexical effect by prompting half to be rewritten so that most individuals will pick the safe option (insurance). The remaining half was prompted without an added condition since prompting to increase risk taking resulted in hallucinations and low-quality rewrites. We then sourced risk taking behavior data via Prolific Academic, introducing an attention check question to enter the study, followed by allocating each participant a random rewrite to answer, for an average of 200 choices per isomorph. Participants stated their age and gender and were debriefed.

To obtain reliable estimates of mean risk-taking behavior for each ChatGPT-written version, we recruited 3,998 participants via Prolific Academic to provide their choices. After dropping individuals who took the survey twice we were left with 3,992 participants (M_{age}=37.3, 49.7% male, 49.2% female, 1.1% other/prefer not to answer). After

applying our AI attribution pipeline and rewriting the problem in risk-inducing and risk-inhibiting variants, we recruited 1,190 participants via Prolific Academic (M_{age}=38.7, 49.3% men, 50.3% women, 0.4% other/prefer not to answer) to collect preference data (pre-registration here: https://tinyurl.com/muw5vmzs). For our investigation of whether the pipeline is feasible to execute using the smaller-scale model (SBERT), we recruited an additional cohort of 1,197 participants via Prolific Academic to collect risk preferences for rewrites crafted using attributions sourced from the smaller model (M_{age}=39.3, 49.7% men, 49.7% women, 0.6% other/prefer not to answer; pre-registration here: https://tinyurl.com/y8mj828e).

To demonstrate the applicability of the method to out-of-sample instances, we applied our pipeline to this data and sourced phrase-based attributions to craft new risk-inducing and risk-inhibiting versions twice, once only using the isomorphs that the model has been exposed to (training subset), and once using a subset of held-out isomorphs that the model was not exposed to nor selected against (test subset). Thus, we obtain four new variants: training-data based risk inducing, training-data based risk inhibiting, test-data based risk inducing, test-data based risk inhibiting. We used Prolific Academic to collect preference data on the attribution-based risk-inducing and risk-inhibiting rewrites, randomly assigning participants to a 2 (training-data based vs. test-data based) x 2 (risk-inhibiting vs. risk-inducing) full-factorial design. Participants answered an attention-check question, observed one rewrite and answered with their preferred alternative, stated their age and gender, and were debriefed.

To provide researchers initial insights into the role of the chosen neural network for embedding texts we repeated the above analysis using the encoder-only SBERT (Reimers & Gurevych, 2018) instead of the decoder-only Llama, with the same train/test splits and model fitting procedure, followed by applying the same approach and parameters for estimating phrase-based attributions setting the reference token to be SBERT's end-of-sequence token as the two models differ in their tokenization and the chosen token is closest to Llama's end-of-message token in its functionality. We then crafted risk-inducing and risk-inhibiting versions and sourced risk preferences on Prolific Academic by randomly assigning participants to a 2 (training-data based vs. test-data based) x 2 (risk-inhibiting vs. risk-inducing) full-factorial design. We ensured participant validity as before by using an attention check, upon whose completion participants observed a randomly-allocated rewrites and made their preferred choice before reporting their age and gender and being debriefed.

Results

Study 1

Study 1 tested whether individuals are sensitive to naturalistic variations in the language used to describe the same decision task, and whether any such sensitivity can be effectively explained using features of the language.

We find clear evidence that individuals' risk preferences are sensitive to aspects of language other than the frame. Risk taking varied across different rewritings of the problems with the same probabilities and outcomes expressed in the same frame (Figure 1), with the variance in responses significantly exceeding that expected from sampling variation for nearly half of the seed frames. An example of these linguistic effects is a sensitivity to communication styles. We find that ChatGPT and human authors varied in their rhetorical styles, which elicited greater risk taking in isomorphs authored by ChatGPT relative to human rewrites, despite not instructing ChatGPT to increase risk taking ($\beta = 0.082$, SE = 0.038, $p = .032$). In particular, ChatGPT adopted a greater informational focus and more impersonal descriptions compared to the human authors, who tended to generate rewrites aligning more closely with involved and interpersonal communication styles (Biber & Gray, 2013). When accounting for these stylistic differences, ChatGPT rewritten isomorphs were no longer significantly associated with greater risk taking.

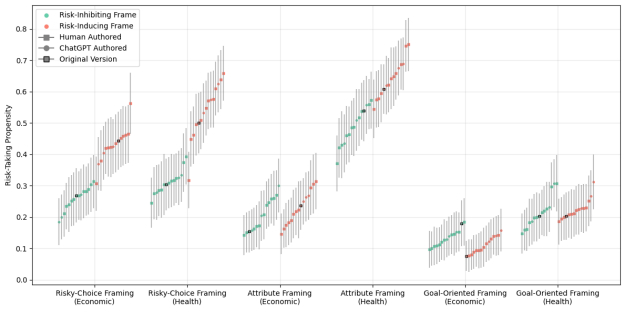

Figure 1: Variability in risk taking across problems and their isomorphs.

We next represented each isomorph as a high-dimensional vector ('embedding' hereafter) using Llama and tested whether introducing the components of these embeddings to a regression predicting individual risk taking from classical covariates added significant explanatory power. We found that accounting for language statistically significantly increases model fit when compared to the baseline psychological model that included terms for problem type, domain, frame type, and individual differences ($\chi 2(29) = 389.0$, $p < .001$), further supporting the idea that there is meaningful signal in seemingly arbitrary linguistic variation.

The presence of additional linguistic predictors of risk taking invites the possibility that framing effects—which generally vary a single dimension of language at a time—may be contingent on such predictors' presence or absence, and so may either be less robust or less practically meaningful when varying many dimensions of language simultaneously. To investigate this possibility, we introduce the notion of "semantic robustness". In this context, semantic robustness captures the extent to which a framing effect is contingent on a specific wording of the two frames of the problem versus generalizes across many different wordings. Intuitively, this would be demonstrated in Figure 1 by separation between the overall sets of risk-inhibiting and risk-inducing frames, rather

than merely a difference between the specific published stimuli (captured by black outlines in the figure). We find that this semantic robustness varies considerably across choice problems. In particular, the economic risky choice problems exhibited the greatest robustness and health goal-oriented framing problems exhibited the least robustness. These results highlight that the degree of stability of effects does not depend only on sampling stimuli from diverse participants and domains, but also on different wordings of what might seem like arbitrary aspects of textual stimuli, which can—and do—impact results.

Study 2

Study 2 tested whether explainable AI techniques can reliably detect the contextual influences of linguistic elements on risk taking and whether these insights can then be used to manipulate behavior.

Applying our analytical pipeline to Study 1's data outputted a set of attributions that rank words and phrases by their association with risk taking (where higher rankings suggest relatively greater association with risky choices and vice versa). In Figure 2, we visualize these attributions, mean centering and scaling their respective values within seed frame to unit variance and positioning them based on the value of their associated texts' first principal component, the latter capturing the variation in semantics, context, and the number of characters among words and phrases.

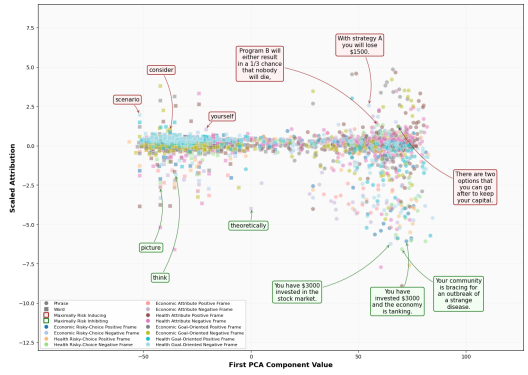

Figure 2: Scatter plot for attributions sourced from each seed frame

We observe that phrases tend to generate more varied attributions with larger magnitudes than words do. To validate these attributions, we used them to redesign each seed frame in two variants: one risk-inducing version using the lexical items attributed as contributing most to risk taking and one risk-inhibiting version using those attributed as contributing most to risk avoidance. We then recruited 971 new participants who provided a total of 11,652 choices across these rewrites to test the efficacy of our method. We find that the risk-inducing isomorphs generated a 9.51% relative increase in the mean risk propensity of participants compared to their risk-inhibiting variants ($\chi2(1) = 11.614$, $p < .001$). Furthermore, our manipulations were successful for both word ($\beta = 0.168$, $SE = 0.075$, $p = .026$) and phrase-based rewrites ($\beta = 0.18$, $SE = 0.072$, $p = .011$) in analyses accounting for heterogeneity by using clustered standard errors at the participant level, participant fixed effects and stimulus fixed effects/

We find that while the test for heterogeneity is non-significant for phrase-based rewrites ($p = .091$), for word-based rewrites it is significant ($p = .03$). To attempt to explain any heterogeneity in the effectiveness of our method, we conducted further analyses using structural equation modelling to predict the success of these manipulations. We find that better results are obtained with models that can more accurately predict behavior on rewrites that are sufficiently similar to the training data. When these factors are aligned, the predicted risk propensities on the risk-inducing and risk-inhibiting rewrites are better calibrated, which minimizes the error of the predicted risk separation effect. Likewise, regularization strength explains the heterogeneity of the word-based condition by reducing sensitivity to word-based but not to phrase-based alterations. In short, our method is more likely to succeed in cases where the linguistic signal is sufficiently strong to facilitate prediction, and when rewrites do not excessively alter the structural and lexical characteristics of the textual stimuli.

Study 3

Study 3 further applies our method to a real-world decision-making task where respondents decide whether to buy optional insurance for an upcoming vacation. Here, we found that prompting ChatGPT-4o to induce greater risk inhibition in rewriting this problem was successful (17.85% relative increase in the relatively risk-inducing isomorphs, $\chi2(1) = 13.154$, $p < .001$).

We then applied our analytical pipeline to this new task's data twice, first designing rewrites crafted from data the model has seen in training and then, separately, using a held-out test subset. Investigating whether these rewrites successfully manipulated risk taking served both to replicate earlier results in a context under ambiguity with reduced model memorization concerns, while also testing its potential to design stronger interventions when more signal for risk taking was present in the training data. Notably, we can compare the efficacy of our method to ChatGPT-4o's naïve manipulation of risk taking. Examining the resultant choices of 1,190 participants, we find that both training (61.06% relative increase, $\chi2(1) = 14.532$, $p < .001$) and test (36.06% relative increase, $\chi2(1) = 7.991$, $p = .005$) conditions significantly manipulated risk taking, with no significant differences in their effects ($\beta = 0.202$, $SE = 0.248$, $p = .416$). Furthermore, our approach resulted in larger effects than those obtained by prompting ChatGPT to induce such manipulation (interaction term for the approach's risk manipulation exceeding that of ChatGPT: $\beta = 0.344$, $SE = 0.141$, $p = .014$).

Importantly, rich linguistic feature representation made recently possible with the advent of LLMs, is critical to the success of the method, as shown by the fact that replicating this pipeline but replacing Llama with the earlier generation,

smaller scale model SBERT does not result in statistically-significant risk separation in either the training ($\chi2(1) = 2.133$, $p = .144$) or test ($\chi2(1) = 0.801$, $p = .371$) conditions, across 1,197 participant choices. We find evidence that earlier generation models do not extract genuine risk-manipulating linguistic features but rather memorize the training data, diverging significantly from features that LLMs are now able to represent. Interestingly, these insights dovetail with neuroscience research suggesting that improved language-generation capabilities also enable better representations of language-correlated brain activity (Schrimpf et al., 2021).

Finally, we demonstrate our method's potential for inductively mining new hypotheses to test from the data that can generate new theories. For this analysis, we focus on interpreting the phrase-based attributions extracted in Study 3, as the greater contextual information encoded in these is likely to be more practically meaningful. Three mechanisms in particular stand out. The first reframes the vacation as a 'dream vacation', which could be inducing a reverse endowment effect increasing individuals' willingness to pay, and fits with emotional positioning used in marketing to elicit a connection by appealing to emotion over logic (Mahajan & Wind, 2002). Two additional coherent mechanisms that to our knowledge have not been covered in extant literature are price minimization and a concept that we term 'questions of worth'. The former complements the vacation's pricing by a minimizing adjective (e.g., "for just $60…"), and the latter reframes the decision task not as a question of desire ("do you want…") but as a question of worth ("is it worth…"). This could be shifting the focus from an actual spending decision (i.e., following through with the purchase) to a hypothetical cost-benefit analysis, which further supports buying the insurance. Though not a conclusive test, we find that there are statistically-significant risk-inhibiting effects to price minimization ($\chi2(1) = 10.263$, $p = .001$), questions of worth ($\chi2(1) = 17.923$, $p < .001$), and emotional positioning ($\chi2(1) = 14.702$, $p < .001$). Although these insights are exploratory and meant to serve as an existence proof, they stress the potential for the method to be used for new theory generation by freely-identifying linguistic triggers of human behavior.

Discussion

Taken together, the present research makes important contributions at three levels: At a theoretical level, we reveal what elements of language drive behavior, in particular surrounding risk, both recovering previous theories and forming new hypotheses freely from the data. We also demonstrate that the semantic robustness of psychological effects to linguistic alterations varies considerably across effects, providing important context with which to interpret extant theories. At a methodological level, we introduce a general method for tracking interpretable relationships between language and behavior that can be widely applied across areas of social science to understand how elements in textual stimuli shape behaviors. Finally, practically, our method can both contribute to accelerating scientific research via automating the process by which behaviorally important features of language are discovered, as well as streamlining organizations' and governments' efforts to craft communications by identifying optimal language use in different contexts efficiently from less data.

There are myriad future directions that scholars could investigate. Given limited overlaps in the vocabularies across seed problems, we have not examined how the attributions of specific words or phrases change as context changes. For example, how does the relationship of a focal phrase with risk taking change based on the domain in which it is used? Through our method, we can test and quantify such foundational aspects of the contextual nature of meaning as have been theorized by linguists and philosophers such as Wittgenstein, as the same wording may realize different meanings in different contexts (Halliday, 1978). Moreover, we can use our method to gather process evidence regarding how other manipulations (e.g., the modality of presenting text data, as written or spoken) change patterns of activations to further understand mechanisms of attention. On the technical side, one could further develop a method for automatically extracting psychological theories from the patterns of generated attributions, for example by relying on the concept of LLM-as-a-Judge, which has been shown to be a powerful and scalable tool for uncovering human preferences (Zheng et al., 2023), as well as further replication studies using different LLMs to understand exactly what dimensions of meaning it is necessary to encode to replicate our effects.

Some important limitations to be cognizant of include our scoping results regarding the drivers of prediction errors discussed in Study 2, that give conditions under which it is more likely to be able to extract meaningful predictors of behavior. Likewise, as our analyses recruited a US-based cohort of English-speaking decision-makers, the method's ability to detect linguistic influences applicable across cultures and geographies remains to be tested. Furthermore, decision-making under risk and ambiguity is a task that is sensitive to affect (Loewenstein et al., 2001); it could be that effects would be weaker for less emotionally laden tasks. Lastly, our corpora introduced isomorphs with a few hundred tokens at most; longer passages of text may necessitate advances in LLMs' context window sizes to extract meaningful features.

Our findings show that LLMs and explainable AI can be harnessed to identify how language contextually affects human behavior both efficiently and at scale. With this understanding, it is possible to design interventions that effectively manipulate behavior—for example, discouraging individuals from taking undesirable health risks—while simultaneously facilitating building new, general theories from the uncovered patterns of associations. Though we have focused on the domain of risk, our approach can be broadly applied to understand the relationship between language and behavior across contexts, unlocking previously hidden insights with immense potential for aiding and understanding human decision making and behavior more broadly.

Acknowledgments

The authors disclose the use of AI software (ChatGPT, Claude) to generate some portions of code used in the analysis.

References

Bhatia, S. (2019). Predicting risk perception: New insights from data science. Management Science, 65(8), 3800-3823.

Bhatia, S. (2024). Exploring variability in risk taking with large language models. Journal of Experimental Psychology: General, 153(7), 1838.

Biber, D., & Gray, B. (2013). Identifying multi-dimensional patterns of variation across registers. In Research methods in language variation and change (pp. 402-420). Cambridge University Press.

Boselli, R., D'Amico, S., & Nobani, N. (2025). eXplainable AI for word embeddings: A survey. Cognitive Computation, 17(1), 19.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic perspectives, 19(4), 25-42.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Vasic, P. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Halliday, M. A. (1978). Language as social semiotic (p. 136). Arnold: London.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896.

Lawson, M. A., Larrick, R. P., & Soll, J. B. (2024). The Individual Overconfidence Test: A parsimonious measure of trait-level overconfidence. Available at SSRN 4893481.

LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. Journal of Behavioral Decision Making, 16(2), 77-92.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. Organizational behavior and human decision processes, 76(2), 149-188.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. Psychological bulletin, 127(2), 267.

Mahajan, V., & Wind, Y. (2002). Got emotional product positioning?. Marketing management, 11(3), 36.

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. Scientific Reports, 14(1), 4692.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Molins, F., Gil-Gómez, J. A., Serrano, M. Á., & Mesa-Gresa, P. (2024). An ecological assessment of decision-making under risk and ambiguity through the virtual serious game Kalliste Decision Task. Scientific Reports, 14(1), 13144.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2025). A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 16(5), 1-72.

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In Computing attitude and affect in text: Theory and applications (pp. 1-10). Dordrecht: Springer Netherlands.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118.

Şenel, L. K., Utlu, I., Yücesoy, V., Koc, A., & Cukur, T. (2018). Semantic structure and interpretability of word embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1769-1779.

Simonsohn, U., Montealegre, A., & Evangelidis, I. (2025). Stimulus sampling reimagined: Designing experiments with mix-and-match, analyzing results with stimulus plots. Journal of Personality and Social Psychology.

Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.

Tao, C., Shen, T., Gao, S., Zhang, J., Li, Z., Hua, K., ... & Ma, S. (2024). Llms are also effective embedding models: An in-depth overview. arXiv preprint arXiv:2412.12591.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. Journal of behavioral decision making, 15(4), 263-290.

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46595-46623.