**Computational Analysis of 100K Choice Dilemmas**

Sudeep Bhatia
University of Pennsylvania

Simon van Baal
University of Leeds

Feiyi Wang
University of Pennsylvania

Lukasz Walasek
University of Warwick

February 26, 2025

Send correspondence to Sudeep Bhatia, Department of Psychology, University of Pennsylvania, Philadelphia, PA. Email: bhatiasu@sas.upenn.edu. Funding was received from the National Science Foundation grant SES-1847794. The authors do not have any competing interests.

## Abstract

We present a dataset of over 100K textual descriptions of real-life choice dilemmas, obtained from social media posts and large-scale survey data. Using large language models (LLMs), we extract hundreds of choice attributes at play in these dilemmas and map them onto a common representational space. This representation allows us to quantify the broader themes and specific tradeoffs inherent in life choices and analyze how they vary across different contexts. We also present our dilemmas to human participants and find that our LLM pipeline, when combined with established decision models, accurately predicts people's choices, outperforming models based on unstructured textual content, demographics, and personality. In this way, our research provides new insights into the attributes, outcomes, and goals that underpin life choices, and shows how large-scale LLM-based structure extraction can be used, in combination with existing scientific theory, to study complex real-world human behavior.

**Keywords**: decision making, computational modeling, large language models, text analysis, multi-attribute choice

**Significance Statement**: Understanding how and why people make the decisions they do is core to behavioral science, yet applying established theories to important real-world choices poses several technical and theoretical challenges. The emergence of large language models and the surge in publicly available user-generated text data allows us to collect, extract structure from, and analyze, people's day-to-day experiences. We leverage these developments to code the attributes and reasons at play in important real-world choices and predict new choices using a popular decision-making model. This pipeline showcases a novel technique to understand and predict decisions outside the laboratory, furthering our understanding of human choice, and opening new opportunities to improve people's wellbeing.

## Introduction

One of the main goals of behavioral science is to understand how people make decisions and to predict what they choose. To this end, fields like psychology, economics, sociology, business, and neuroscience have developed a wide range of theories that identify the outcomes and goals that people prioritize, as well as the psychological mechanisms and decision strategies they use to obtain these outcomes and goals (1–7). The hope is that by understanding and predicting decision making, we can improve people's choices and in turn enhance their well-being, with positive outcomes for society (8–11).

Yet, despite substantial progress, the quantitative modeling of decision processes has largely been confined to highly stylized artificial stimuli involving just two or three attributes, like the payoffs and probabilities of simple monetary gambles or quality ratings and prices of hypothetical consumer goods. Although these stimuli enable precise quantification, they fail to capture the complex considerations that drive important life choices. Moreover, the way choices between these stimuli are elicited—typically on computer screens in the laboratory, with keyboard presses to indicate choice, and several trials over a short experimental time frame—remains far removed from the embedded situational contexts, interpersonal dynamics, and extended timescales that underlie major real-world decisions.

For this reason, decision science is still struggling to accurately predict and influence real-world choice behavior. For example, gamble choice tasks, a staple in the decision scientists' toolbox, show weak correlations with a willingness to engage in common risky behaviors (12, 13), and are only modestly correlated with other artificial lab-based measures (14). Similarly, paradigms designed to elicit people's time preferences show only small correlations with clinical, financial, and health-related intertemporal behaviors (15–19). Recent analyses of field studies of behavioral interventions have likewise found small effect sizes (20, 21). Reflecting on these challenges, several researchers began to call into question whether real-world decision making can be studied using established empirical and analytical methods (22–26). Any attempt at solving this problem must first find a way to uncover the rich representations that underpin typical decisions people face in their lives.

Two recent cultural and technological developments may help solve this external validity challenge. First, the growth of social media over the past 20 years has created an unprecedented record of people's experiences, typically in the form of textual data. Platforms like Twitter and Reddit capture millions of naturalistic first-person accounts detailing diverse situations and choice dilemmas. This massive volume of data provides a unique window into the psychology of real-life choice (27–31). Second, advances in natural language processing, specifically large language models (LLMs) and generative artificial intelligence (AI), now enable the extraction of structured, quantitative information from textual data (32–34). Whereas in the past, making sense of extensive textual corpora required laborious human analysis, contemporary LLMs can automatically identify the entities, features, and relations present in open-ended narratives, and do so at scale, in a way that can mimic human representations of those entities, features and relations (35–39).

The goal of this paper is to use these cultural and technological developments to model the complex landscape of important life choices. We do so by applying an LLM-based filtering and structure extraction pipeline to millions of Reddit posts about the decisions that people are facing in their own lives. This allows us to build a repository of real-life binary choice problems –

decisions involving precisely two options. We supplement this dataset with additional choice problems from a large-scale survey of US participants with demographic characteristics (age, gender, ethnicity) representative of the US population. We also use our pipeline to extract, for each of the dilemmas in our dataset, a set of natural language reasons describing the costs and benefits involved in the dilemma. Finally, we quantify each of the reasons in terms of the degree to which they reflect various attributes, outcomes, and goals. The attributes considered in our analysis are taken from prior literature, and include a diverse array of personal, romantic, familial, professional, moral, cultural, spiritual, intellectual, emotional, and decision-theoretic considerations (40–51), allowing us to represent our choice dilemmas in terms of theoretical constructs studied by psychologists over several decades.

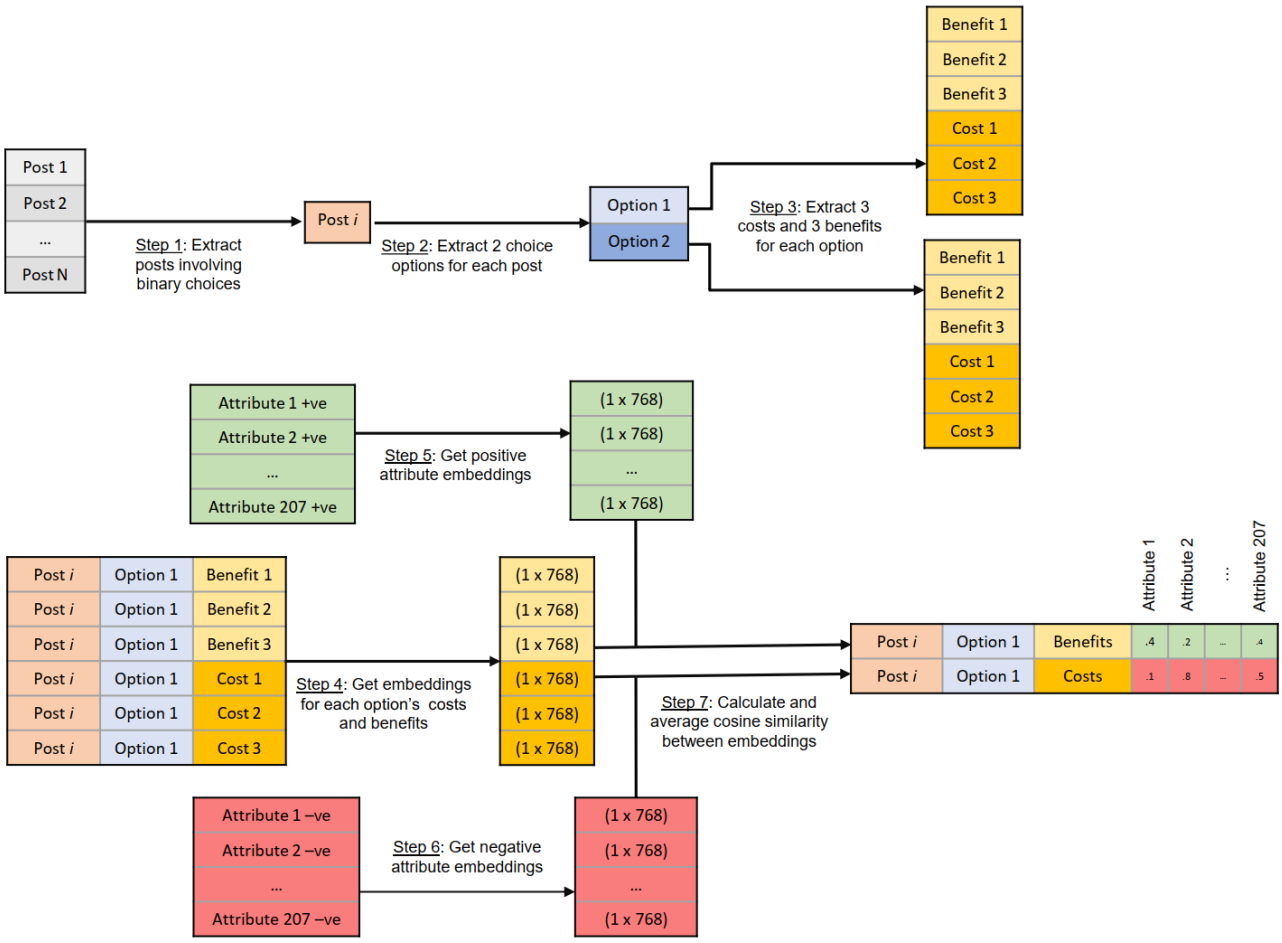


**Figure 1.** The LLM structure extraction pipeline. The top half of the figure represents the filtering and coding process. Here, the LLM identifies posts with binary choices (step 1), and, for these posts (in this example, post *i*), generates natural language descriptions of the two choice options (step 2) and of their costs and benefits (step 3). The bottom half of the figure depicts the transformation of option 1 of post *i* into 207 quantified attribute values. This is done by first obtaining 768-dimensional SBERT embeddings for GPT-generated costs and benefits (step 4) as well as for sentences describing positive and negative examples of the attributes (steps 5 and 6). Finally, the pipeline calculates the average cosine similarity of each positive and negative attribute embedding with each benefit and cost respectively (step 7). This pipeline (with the exclusion of steps 1 and 2) is also applied to everyday decision problems generated by a sample of US participants in Study 1.

**Table 1**: Summary of theory-driven attributes, outcomes, and goals organized in this paper. Here # indicates the number of distinct attributes in each category, and the remaining columns show one example attribute and one example sentence used to code that attribute. There are a total of 207 attributes across all categories, each with several positive and negative sentences, which we use to code the attributes.

| Category | # | Example Attribute | Example Attribute Sentence |
|---|---|---|---|
| decision theory | 5 | safety vs. risk | *prevents risk* |
| decision outcomes | 11 | money | *increases wealth* |
| consumer behavior (42, 52) | 6 | experiential consumption | *produces pleasant memories* |
| emotions (43, 53) | 6 | anger | *defuses hostility* |
| self-determination theory (47) | 3 | social relatedness | *fosters a sense of belonging* |
| values (44) | 10 | tradition | *encourages preservation of cultural heritage and customs* |
| goals (40) | 135 | sexual desirability | *enhances romantic appeal* |
| moral foundations (48) | 5 | harm | *helps someone weak or vulnerable* |
| person perception (45, 54) | 10 | competence | *involves being competent* |
| social-value orientation (41) | 4 | competitiveness | *enhances personal status* |
| altruism (50, 51) | 9 | efficiency | *maximizes the use of available resources* |
| fairness (46, 49) | 3 | procedural fairness | *advances fair and inclusive decision making* |

Overall, our LLM pipeline results in over 100K unique choice dilemmas, 200K distinct choice options, and 1.2M natural language costs and benefits. These are coded on more than 200 theory-derived attributes resulting in more than 100M quantified attribute values (a dataset that would simply be impossible to collect with human coders). We validate our coding scheme in two studies (Study 2 and 3) and then use the covariance structure of this dataset to derive a hierarchical taxonomy of the diverse considerations at play in naturalistic decision making, which gives us new insights into the prominence of different decision attributes, their co-occurrence relationships and tradeoffs, and their distribution across different demographic groups and social contexts. Finally, we use this taxonomy, in combination with existing decision models (55–58), to predict people's choices (Studies 4a and 4b) and verbalized reasoning (Studies 5a and 5b) in the dilemmas.

## Results

## Attribute Extraction and Validation

Our primary analysis relied on the r/Advice subreddit, a popular forum on Reddit in which users ask for advice on various everyday dilemmas they are facing. As of February 2025, this subreddit had 1.3M members. For additional tests, we also coded four smaller advice subreddits with a narrower focus (r/careeradvice, r/FriendshipAdvice, r/AskMenAdvice and r/askwomenadvice), and additionally elicited choice dilemmas from a sample of US participants with demographic characteristics representative of the population along the dimensions of age, gender, and ethnicity, in a preregistered study (Study 1). The number of posts and requests for advice in each of these datasets is summarized in Figure 2A. Figure 2B shows the total number of posts in the subreddits that were used in our analysis (the remaining posts were either too long, too short, or did not involve binary choices as assessed by OpenAI's Generative Pre-trained Transformer-3.5-turbo model (GPT)). Figure 2C shows the distribution of the choice dilemmas over time. See Methods for further details of the datasets.

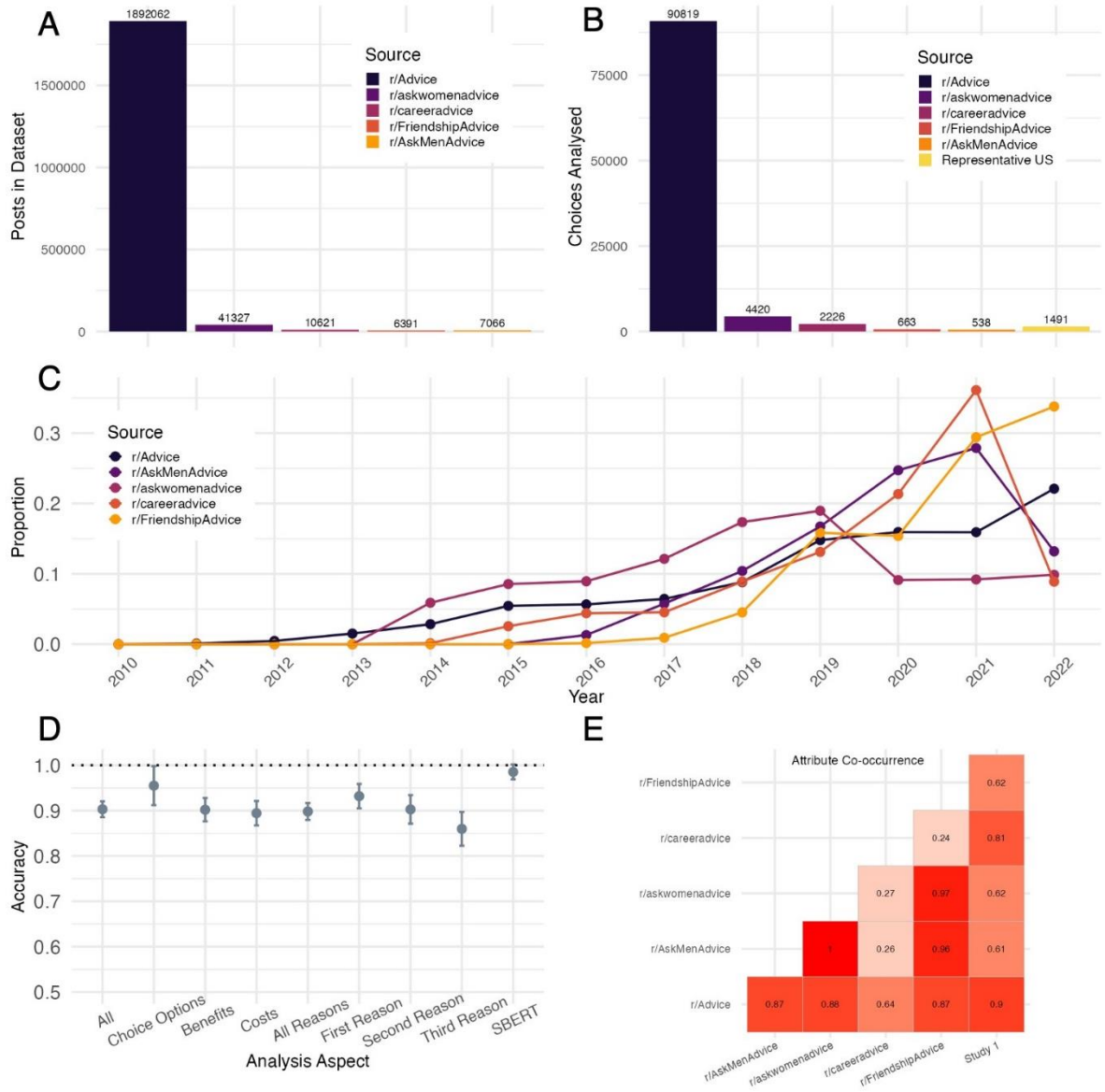


**Figure 2.** Descriptive statistics on the data sources used. (A) and (B) show the number of posts in each subreddit and how many of them were finally analyzed, respectively. (C) shows the post volume of the different subreddits over the years. (D) summarizes participant assessments of the accuracy of LLM outputs. (E) displays the correlation of attributes across the data sources in this paper.

We began our analysis by testing whether our LLM-based computational pipeline could accurately extract the structure and content of the choice dilemmas. Our pipeline had two main components, as shown in Figure 1: First, we used GPT to generate natural language descriptions of the two choice options in each dilemma, along with costs and benefits for each option (34). Second, we used an SBERT sentence embedding model to map these extracted costs and benefits onto 207 decision attributes, outcomes, and goals previously identified as decision-relevant in behavioral science research (59). Our LLM pipeline is illustrated in Figure 1, Figure S1 shows an example of

a post as well GPT outputs and our attribute analysis, Table 1 presents a summary of the 207 theory-driven attributes, and the full set of attributes is presented in Table S3. See Methods for further details of our LLM pipeline.

We validated the first component of the LLM pipeline in Study 2 by asking human participants whether GPT's outputs were accurate for a subset of 100 randomly selected dilemmas from r/Advice. These judgments are shown in Figure 2D. Here we can see that 95.51% of GPT's generated choice options were judged to be accurate by participants, and that the average accuracy of the associated reasons (costs or benefits) was 90.29%. Accuracy was higher for the first reason generated by GPT, and then dropped slightly for the second and third reason, though in all cases accuracy rates remained higher than 85%.

We next validated the sentence embeddings' coding accuracy in Study 3 by selecting, for each of the 207 attributes, GPT-generated reasons (costs or benefits) that were semantically similar or dissimilar to the attribute in sentence embedding space. We then paired one similar reason and one dissimilar reason, and asked participants to judge which of the two reasons was most reflective of the attribute in question. Participant judgments for these attributes are also shown in Figure 2D, which indicates that our pipeline achieved an accuracy rate of 98.53% in quantifying GPT-generated reasons in terms of theoretically derived attributes. All proportions in Figure 2D are significantly different to 50% ($p < .001$) according to a binomial test. See Methods for experimental details for Studies 2 and 3.

We also validated the content of our main r/Advice dataset by comparing the prominence of attributes in this dataset with those in choice dilemmas generated by a sample of US participants in Study 1. Figure 2E shows that there is a very high correlation ($r = 0.90$, $p < .001$, 95%CI = [0.87,0.92]) in the attribute frequencies of these datasets, indicating that the content of r/Advice closely resembles that of important choice dilemmas elicited in a controlled and demographically representative survey. Figure 2E also shows that the other datasets in our analysis covary on attribute frequency in expected ways, with r/AskMenAdvice and r/askwomenadvice being very correlated to each other and to r/FriendshipAdvice, and all three of these subreddits being weakly correlated to r/careeradvice. The dilemmas in r/Advice and in Study 1 are moderately correlated with these four datasets, indicating that their content captures a balanced mix of dilemma types observed in professional and social domains of life. We explore the content differences between these domains in a subsequent section of this paper.

## Structure and Content of Choice Dilemmas

Having validated the accuracy of our LLM pipeline, we next investigated the latent structure in our dataset. More specifically, we performed clustering analyses to identify the core attribute profiles and tradeoffs that describe the choice dilemmas in the online posts. Recall that for each choice option, our pipeline extracted reasons why a person should (benefit) or should not (cost) choose that option, and coded these reasons on 207 theory-derived attributes. We used hierarchical clustering on the attribute distributions for approximately 180K choice options in the r/Advice subreddit. The result of this is visualized in Figures 3A, which labels attribute clusters according to their dominant themes. The unlabeled attribute clusters are shown in Figure S2. SOM 1.1. explains our attribute clustering methods in detail.

From Figure 3A, we can see that a clear and logical structure emerges, with many attributes cleanly dividing into categories like physical health, social desirability, happiness and pleasure,

and risk and stability. Figure 3A also reveals nuance in how attributes covary with each other. For example, social attributes are split between clusters that emphasize social connections (like having friends), clusters that emphasize social perception (like being admired), clusters that emphasize personal virtues (like being loyal), and clusters that emphasize moral outcomes (like distributive fairness). Similarly, money-related attributes are divided between clusters that emphasize financial prudence and pragmatism (like utilitarian consumption) and clusters that emphasize professional outcomes (like making money).

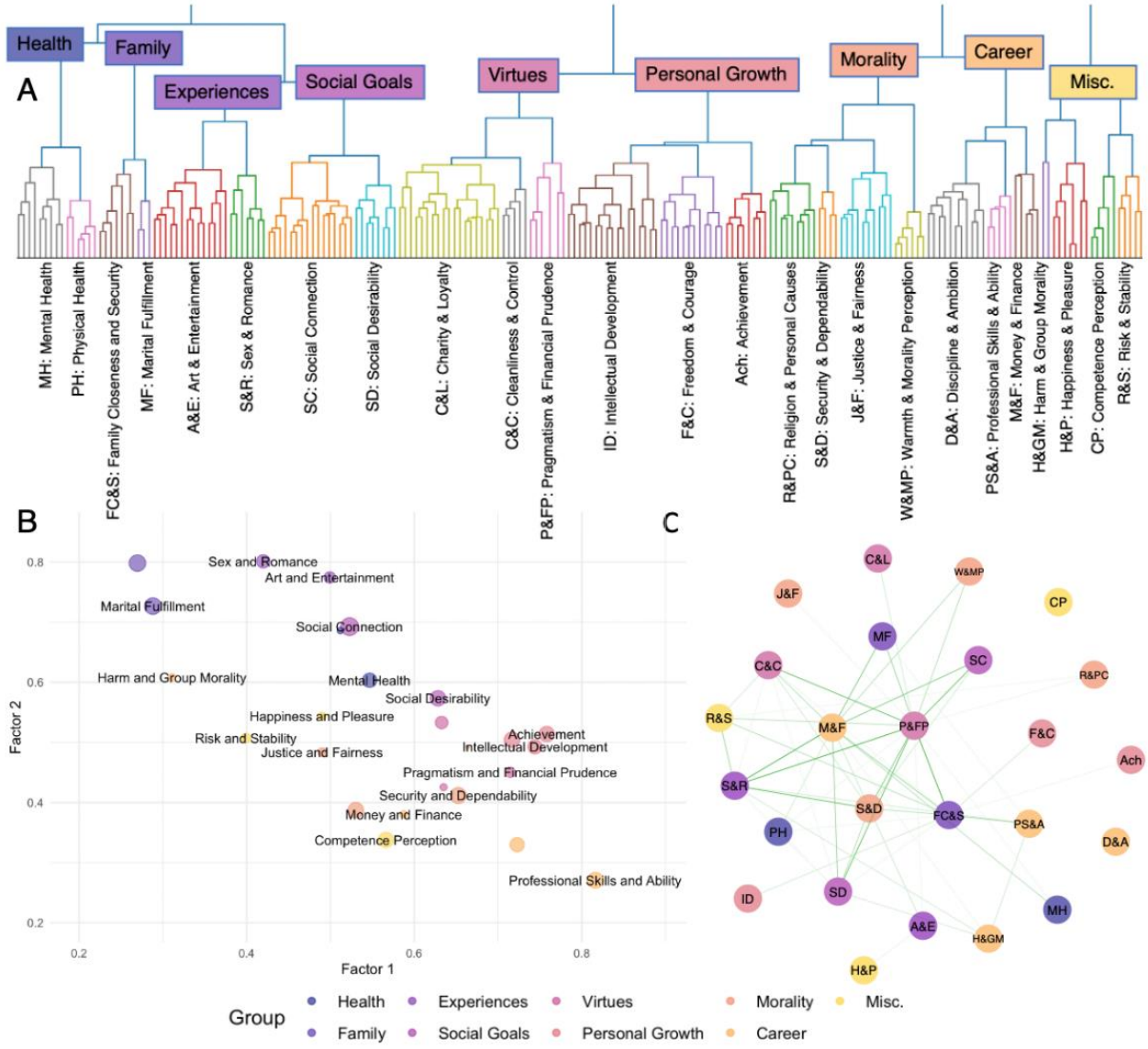


**Figure 3.** The option attribute structure. (A) shows the hierarchical attribute clustering of attribute across options. (B) represents attribute loadings on the two primary factors in a factor analysis. The size of the points in (B) is proportionate to the attribute frequency. (C) shows how often different attributes are present in opposing choice options. The thickness of the lines represents the frequencies of these attribute tradeoffs.

Figure 3B visualizes the structure of the clusters using a factor analysis (see Figure S3 for loadings and SOM 1.2. for details), and shows the types of cross-cluster similarities we would typically expect (e.g. with the money and finance cluster being close to the pragmatism and financial prudence cluster). Figure 3B also displays the average prominence of the attributes in the clusters, indicating that our dataset involves many choices with professional/financial and interpersonal attributes and relatively few choices with religion and physical health attributes. This is why the first two factors revealed by this analysis seem to capture professional/financial and interpersonal/social attributes respectively.

Unlike Figures 3A and B, which are based on how often attributes are likely to co-occur in the same choice option, Figure 3C shows how often attributes occur in separate choice options within the same dilemma. In other words, it displays common attribute tradeoffs. Here we see that many of the choices in our dataset involve tradeoffs between monetary or professional attributes on the one hand (e.g. those in the money and finance cluster or the pragmatism and financial prudence cluster) versus interpersonal or social attributes on the other (e.g. those in the sex and romance cluster, social connections cluster, or the family closeness and security cluster). Monetary attributes also tradeoff against warmth and morality perception, and social attributes (in particular family closeness) tradeoff against attributes involving personal growth (e.g. intellectual development and freedom and courage). We also see interesting patterns involving the risk and stability cluster, which with safety often trading off against both money and finance as well as sex and romance. The pattern of tradeoffs between attribute clusters is shown in greater detail in Figure S4. Also see SOM 1.2. for details of how we calculated attribute tradeoffs.

## Social and Situational Context

One of the strengths of analyzing online forum data is that the data contains rich information about the social settings and time of choice. To leverage this information, we analyzed how the content of choice dilemmas varies across different contexts (see SOM 1.3. for details). For our first analysis, we divided the r/Advice dataset into dilemmas involving a single male, a single female, a same-gender dyad and a different-gender dyad. The prominence of the attributes for each of these groups is shown in Figure 4A. Here we see that posts mentioning only one person (either a man or a woman) are more likely to involve professional, money-related, and personal development-related clusters, with these clusters being slightly more prominent with males than females. Conversely, posts with dyads are much more likely to involve romantic, social, and familial clusters, with different-gender dyads displaying a much greater frequency of romantic concerns than same-gender dyads (see also Figure S5A). Further, we divided the r/Advice data into subsets with a single mature-age (25 or older) individual, a single low-age (younger than 25) individual, as well as dyads with pairs of mature-age, pairs of low-age, and pairs of mixed age (one mature-age and one low-age) individuals. We found that older individuals are more likely to discuss professional and less likely to discuss social concerns than younger individuals. Dilemmas with pairs of young individuals are also much more likely to involve sex and romance, social desirability, mental health, and pleasure concerns, whereas dilemmas with mixed-age dyads are more likely to involve family closeness and security (see Figure S5B). We also explored the prominence of the attribute clusters across relationship categories. The bottom of Figure 4A shows that in r/Advice, the type of relationship dilemma correlates closely with associated attributes detected in the dilemma in our analysis, with, for example, romantic relationships being more likely to involve marital fulfilment and sex and romance clusters, and professional relationships being more likely to involve money and finance and professional skills clusters.

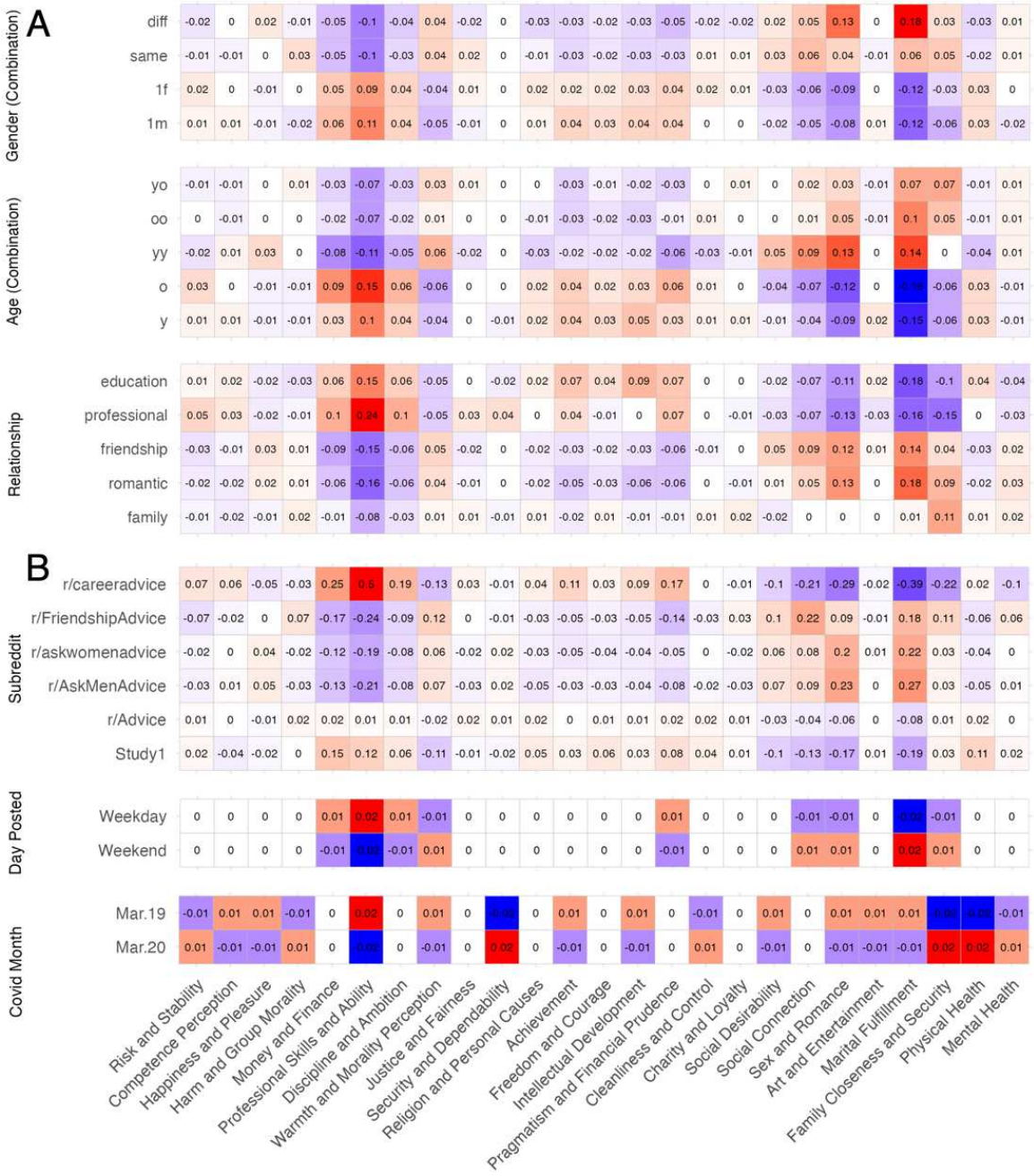


**Figure 4**. Attribute cluster heatmaps. (A) displays the effect of gender composition, age composition and relationship on attribute cluster frequencies. “m” And “f” indicate male and female respectively, and “y” and “o” indicate young and old (threshold of 25 yr) respectively. (B) displays the effect of forum, day of the week, and the COVID pandemic. We mean normalized the cells for each grouping, for each attribute. Thus, for example, for the risk and stability attribute and the gender groupings, we first obtained the extent of the attribute in posts with a single male, a single female, a same gender pair and a different gender pair. Then we mean-normalized this 4-item list, to get the relative extent to which the risk and stability attribute manifests for each of the four gender groupings. This gave us the values in the cells. The shading of each cell is based on the maximum and minimum of each block.

Figure 4B illustrates a similar analysis in which we examine cluster prominence across datasets. Consistent with Figure 2E, r/AskMenAdvice and r/askwomenadvice have very similar cluster profiles to each other and, to a slightly lesser extent r/FriendshipAdvice. All three of these subreddits emphasize social attributes, with sex, romance and marital fulfillment being more common in the first two and social connections and desirability being more common in the third. Conversely, r/careeradvice emphasizes professional, money-related, personal development-related, and risk-related clusters, and deemphasizes social, romantic, and familial clusters. Finally, the contents of r/Advice dilemmas are similar to the dilemmas generated by the US sample in Study 1. The dilemmas involve a balance of professional and social attributes. Of note for Study 1 is that the prevalence of sex and romance-related attributes is less than r/Advice, reflecting that participants may feel less comfortable talking about such topics in a survey.

Our last analysis, shown at the bottom of Figure 4B, examined time effects on the content of the dilemmas. For this purpose, we divided dilemmas in r/Advice based on whether they were posted on a weekday or weekend and whether or not they were posted in March 2020 (at the onset of the COVID epidemic) or in March 2019 (prior to the epidemic, but at the same time of the year). Expectedly, weekday posts typically involve profession and work-related attributes whereas weekend posts typically involve social, romantic, and familial attributes (see also Figure S6A). Additionally, posts made at the onset of the COVID epidemic involve a slightly higher frequency of health, family closeness and stability, cleanliness and control, security and dependability, and harm and ingroup morality, and risk and stability-related attributes (see also Figure S6B).

## Choice Prediction

So far, we have shown that our approach can accurately quantify the attribute compositions of, and tradeoffs inherent in, choice dilemmas. This implies that formal decision theories developed by researchers (theories that describe how people resolve attribute tradeoffs) can be applied alongside our pipeline to predict choices between life dilemmas. One such theory is the weighted additive rule, which proposes that people have attribute weights (with positive weights on an attribute indicating a preference for that attribute, and a negative weight indicating an aversion for that attribute), and that an option's utility is simply the weighted sum of its attributes (55–58). If our computational pipeline accurately codes the attributes in the dilemmas, then our LLM-coded attributes should be able to predict choices when combined with the weighted additive rule.

We tested this in the preregistered Study 4a, which used r/Advice dilemmas that involved the most common attribute tradeoffs in our dataset – the tradeoff between the family closeness and security cluster, and the pragmatism and financial prudence cluster. Study 4a offered participants eight dilemmas with this tradeoff, and asked them to rate their preference for the options in each dilemma on a Likert scale (see Methods for details of the experiment, and SOM 1.4. for details of stimuli generation). Our main goal was to predict each individual's eight preference ratings using an individual-specific (i.e. "within-subject") weighted additive decision model (SOM 2.1.). To minimize overfitting, we used a dimensionality reduced version of our 207-dimensional attribute space. As shown in Figure 5A, this decision model achieved an average $R^2$ of 0.24 across our participants. We compared this model against two alternatives. The first used the unstructured textual content of the posts (without LLM-based attribute extraction) and the second used randomly generated attributes (SOM 2.2.). Both alternates achieved significantly lower $R^2$*s* of 0.14 showing the superior performance of our LLM-derived attributes. Preregistered Study 4b repeated this test with another common attribute tradeoff – the tradeoff between marital fulfillment and

money, and found a nearly identical result. Note that the random model achieves a positive $R^2$ is because each fit involves only eight observations, indicating overfitting. However, our decision model significantly outperforms the random model as assessed by a paired t-test ($p < 0.01$ for both studies; see Table S1 for additional statistics). Also note that all models tested here have an equal number of free parameters.

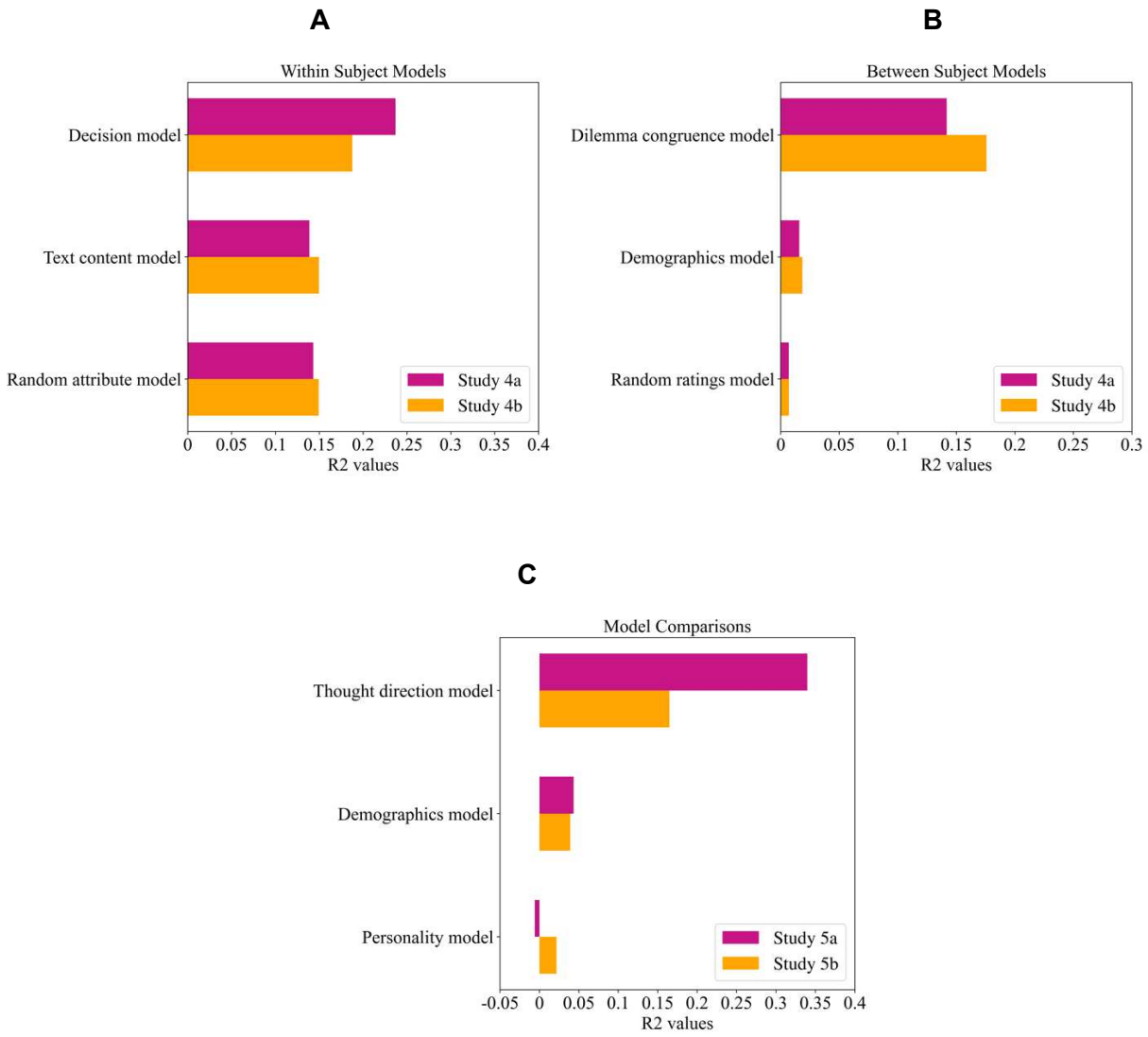


**Figure 5**. Model goodness-of-fit. (A) displays the $R^2$ values of the within-subject models in Studies 4a and 4b. (B) displays the $R^2$ values of the between-subject models in Studies 4a and 4b. (C) displays the adjusted $R^2$ values of the models in Studies 5a and 5b.

The above approach predicts individual-specific preference ratings for the eight choice dilemmas using individual-specific decision models. However, to test the robustness of our approach we also attempted to predict between-participant ratings for each of the dilemmas using preference ratings for other dilemmas that are congruent (involve similar LLM-based attribute structures) or incongruent (involve dissimilar attribute structures) (SOM 2.3.). The key difference between this model and the individual-specific model is that the latter is trained at the individual level and predicts a participant's rating for a given dilemma based on weights on the specific dilemma's

attributes, whereas the between-participant model is trained at the dilemma level and make predictions based on the participant's response to other dilemmas with similar attribute structures. We compared this congruence-based model to a model attempting these predictions with randomly generated attributes (SOM 2.4.), and a demographic model attempting the same predictions using only age and gender. The demographic model is not a standard decision model, but it provides the most appropriate way to compare our approach with demographic predictors, which are, outside the language of the post, the only observable alternative characteristics. We found that the congruence-based decision model achieved an average $R^2$ of 0.14 across the dilemmas in Study 4a and 0.17 in Study 4b, greatly outperforming the predictions of the demographic and random models in the two studies (all $R^2$s < 0.02). These results are shown in Figure 5B. Additional statistical comparisons are provided in Table S1. Note that all models tested here have an equal number of free parameters.

Studies 4a and 4b showed that our approach can be used to extract meaningful information from texts to predict choice outcomes. In preregistered Studies 5a and 5b, we further tested if our approach could describe how people think through these dilemmas. In these studies, we asked participants to list their thoughts (60–62) as they deliberated through decision scenarios taken from Studies 4a and 4b respectively. We found that participants listed thoughts that were highly similar in textual content to the LLM-based attributes of the dilemma they were given (all $p < 0.001$ in Studies 5a and 5b), compared to other dilemmas, indicating that our LLM-based attribute extraction process captures dilemma-specific thought processes in decision making (SOM 2.7.).

We also asked participants to code their generated thoughts based on whether they involved costs or benefits for each of the two options. We found that the direction of people's self-coded thoughts (the difference in the number of benefits vs costs listed for options 1 vs option 2) predicted their preference for option 1 vs. 2 with $R^2 = 0.42$ in Study 5a and 0.32 in Study 5b (SOM 2.5.). We also asked LLMs to code the participant-generated thoughts and found that the direction of LLM-coded thoughts was highly correlated with participants' own coding ($r = 0.66$ in Study 5a and 0.47 in Study 5b). For this reason, we could directly predict people's preference ratings using LLMs applied to their thoughts, achieving $R^2 = 0.34$ in Study 5a and 0.17 in Study 5b. We compared our LLM approach to alternate models relying on eight demographic variables and five personality dimensions(63) (SOM 2.6.). Unlike previous tests in Study 4a and 4b, these alternate models had many more predictors than our LLM approach, which used only a single thought direction variable. Nonetheless, our LLM approach achieved both a higher adj. $R^2$ and a higher raw $R^2$ than the demographics model and the personality models in Study 5a and 5b. Adjusted $R^2$s are shown in Figure 5C and both $R^2$s are shown in Table S2.

## Discussion

Throughout our lives, we face multiple difficult decisions involving tradeoffs between distinct values and goals. Yet, it is hard to imagine that any two decisions are the same, as any one likely reflects people's heterogeneous experiences, contexts, and preferences. This complexity inherent in everyday choice poses a serious challenge to scholars who study decision making, especially those who wish to describe decision making with formal computational or statistical models.

Here we address this challenge by presenting the most comprehensive analysis of real-life decision problems thus far. Using large-scale digital datasets as well as a new LLM-based

analysis pipeline, we extracted hundreds of attributes at play in over 100K real dilemmas. Our findings demonstrate that despite the diversity of choices, distinct groups of attributes and tradeoffs emerge. Many choices people struggle with involve tradeoffs between financial stability/prudence and various social and personal experiences and pleasures. Closeness to one's family is also often contrasted with choices that bring achievement, freedom, and intellectual development. Among other tradeoffs, sex and romance-relevant outcomes are compared with outcomes that promote justice and fairness, as well as security and dependability. Our analysis of the demographic profiles of people who face choice dilemmas, and the timing of these dilemmas, also shows that these tradeoffs differ systematically between different social and temporal contexts. Finally, and perhaps most importantly, the results of our experiments demonstrate that the attribute structure obtained in our analysis can be used to predict hypothetical naturalistic decisions and associated thoughts using a well-established quantitative model of decision making. Importantly, this approach outperforms alternate methods from natural language processing and methods relying on demographic and personality data. Future work applying our attribute-based decision model to specific groups and contexts can be used to generate new insights about decision making. For example, we could examine how the prioritization of different attributes changes across life stages, or in response to societal events, or different social situations. This approach can provide insight into the variability of decision-making processes, helping us understand why similar individuals might make different choices in different settings.

One may wonder whether choice dilemmas extracted from a large corpus of social media posts accurately reflect the difficult and important decisions that shape people's lives. We find support for this in Study 1, which shows strong agreement between the choices reported by a demographically representative US sample and the choices mentioned in r/Advice. It is important to note that Reddit (sensibly) restricts the type of choices that we can observe; for example, the moderation on r/Advice asks its users not to post about automotive choices, seek tech support, or request opinions about their appearance or talents (presumably because there are more specialized subreddits devoted to these types of issues). We did not impose any restrictions on the dilemmas listed by our survey respondents, and the fact that we nonetheless observe a high correlation between decision structures on Reddit and our survey assures us that our results capture many important choices in real life.

Of course, our data is concerned with a particular subset of difficult and typically high-stakes choice dilemmas. This tendency towards big, difficult decisions in our data is because many mundane everyday choices, like what to eat for breakfast, are less likely to feature on online advice forums. In fact, the complex and significant life decisions that feature in our dataset may differ from these simpler everyday dilemmas in more than just the types of attributes that they involve. Important life decisions (64, 65) are likely to be rare, highly uncertain, and consist of numerous conflicting cues and values (i.e., involving incommensurable goals, see (25)). Moreover, these challenging decisions often carry costly or irreversible consequences that may even impact a person's sense of self and their identity (66). That said, while our study offers insights into some of the most important life dilemmas, we acknowledge that the broader landscape of everyday decisions also includes more routine types of choices, and we believe such choices could also be studied using our general computational approach.

Another possible deviation from everyday decisions is that we confined the study to choices with two options, which could mean that some possible insights from multi-alternative decisions are

not included. We do not believe this is particularly worrying since the final choice between the two best options is usually of chief interest and tends to mimic real-life decisions. It has also been theorized that evidence accumulation in multi-alternative choices occurs through binary comparison and that similar processes are used for binary and trinary choices (67–69). That said even though we applied our methodology to explore the binary choices that people share and discuss on Reddit, our methods of LLM-based structure extraction could be easily applied multiple-option choices, and could be used to predict choices in these problems using existing multi-alternative decision models (55–58).

Our LLM pipeline relies on people's ability to describe reasons for and against each choice option in Reddit posts. One concern could be that some people who post on Reddit position are simply attempting to seek confirmation for or feedback on the choice that they have already made. While this is a possibility, note that Reddit's r/Advice moderation rules explicitly prohibit posts that ask for validation or affirmation of prior judgments, opinions, or feelings. There are also other subreddits focused entirely on such post-decisional evaluations, and users seeking mere validation are more likely to post in those dedicated spaces rather than r/Advice. It is also possible that dilemmas on r/Advice are described in a way to make the poster appear desirable to others. Such self-presentation effects are common in many social media spaces but are likely to be less prevalent on r/Advice since one of the primary motives for impression management is influencing other's behaviors (70) which is not a goal in advice-seeking contexts. Additionally, the anonymous and depersonalized nature of the platform indicates that public-identity concerns, and motives to please the audience, are less likely to be present (71). Finally, misrepresenting information could lead to suboptimal advice. Thus, there are few benefits and significant costs associated with providing inaccurate descriptions of life dilemmas (72).

The second and related issue concerns people's access to their own mental states and reasons for their behavior (73, 74). In particular, certain attributes influencing decisions may operate outside the decision maker's awareness. For example, implicit biases could shape social decisions without being consciously accessible (75, 76), making them absent from self-reports on Reddit. Likewise, some attributes are not easily verbalizable. Thus, even if the decision maker is aware, they may struggle to articulate visual or sensory factors that significantly affect decisions. This limitation suggests that the attribute structures revealed by our analysis may be incomplete (77), and that uncovering non-verbalizable or implicit influences remains a challenge controlled lab methods may address more effectively. Nonetheless, we believe our work remains valuable because it captures the deliberative contents of decision-making processes on a large scale and demonstrates (in Studies 4 and 5) that these contents predict choice outcomes. This aligns with a long tradition of research using process-tracing methods, such as reason listing, to model choice processes during decision-making (60–62)·(78, 79)·(80).

Due to its success at uncovering the contents of deliberative thought, we believe that the type of LLM pipeline used in this paper can be used to solve many other types of research problems as well. For example, LLMs can code moral considerations in legal proceedings, public debates over political or ideological issues, and attribute preferences in consumer reviews. Extracting structure from the vast amount of natural language data on these topics will enhance our understanding of how people find and process information, and ultimately make decisions. LLMs are reshaping science, and we believe that our paper illustrates one way this new technology can advance core research in the behavioral sciences.

## Methods and Materials

### Data Availability

All data, stimuli, and code, for this paper can be found in our OSF repository: https://osf.io/f29be/. Note that in line with data sharing practices for Reddit, we have removed the post content and author name for the post. These can be downloaded separately using the post ID from Reddit's official API.

### Choice Dilemmas Datasets

**Reddit.** We obtained Reddit choice dilemmas from monthly data dumps collected by pushshift.io, a Reddit API service managed by the non-profit Network Contagion Research Institute. These datasets contained posts made till the end of December 2022. The total number of posts in these datasets is shown in Figure 2A. We removed all posts that had fewer than 500 characters and more than 2,500 characters before passing them through our LLM coding pipeline, described below. The lower bound was selected so that posts had enough information about the choice dilemma to enable our analysis. The upper bound was selected so that the LLMs could hold all the information in the context to analyze the post. Long posts would have also greatly increased our API costs.

**Study 1.** We obtained choice dilemmas in Study 1 through Prolific Academic. Our sample was representative of the US population stratified across three demographics: age, sex, and ethnicity. We recruited a total of 500 participants (48% male, mean age = 48), who were each asked to describe three important dilemmas involving choices between two options that they have faced in their life. For each dilemma participants were first asked to think about the dilemma, then asked to briefly describe the two choice options they faced, and then finally asked to describe the dilemmas in detail, with emphasis on the aspects and attributes of the choice options, as well as the tradeoffs involved and the outcomes and goals that the participant prioritized. Descriptions were constrained to be at least 250 characters in length. Our experimental interface prevented participants from copying the experimental prompt and pasting pre-generated text into the response box, minimizing the use of generative AI. Note that three participants timed out before submitting their response, resulting in a total of 497 participants who produced a total of 1,491 choice dilemmas that were ultimately passed through our LLM pipeline described below (this minor attrition should not affect our results). Study 1 procedure was preregistered at https://osf.io/hp3sq. Details of experimental materials (e.g. wording of participant questions) are in SOM 3.1.

### LLM Pipeline

**GPT**. We used GPT 3.5-turbo, accessed through OpenAI's API, for the first step in our pipeline. We passed the title and text of each Reddit post to GPT, and asked it to determine whether or not the post involved a choice dilemma with precisely two choice options, and, if so, describe the choice options using a short phrase. After obtaining textual descriptions of the choice options, we asked GPT to generate six short sentences describing three benefits and three costs for each of the two options, described in the text. We used an identical pipeline for the participant-generated dilemmas in Study 1, except that we did not ask GPT to determine whether or not the dilemmas had a binary choice or describe the choice options in the dilemmas (since all dilemmas in Study

1 had already been constrained to be binary choices and involved participant-generated descriptions of these choices). The complete script for querying GPT is provided in SOM 4.

**Sentence Embeddings.** The above step resulted in a total of twelve sentences describing the reasons (costs and benefits) inherent in each choice dilemma. We coded these sentences in terms of the 207 attributes in Table 1 and Table S3 with sentence embeddings. As shown in these tables, we had specified, for each attribute, a list of ten sentences describing positive (or beneficial) instances of that attribute and ten sentences describing negative (or costly) instances of that attribute (for the 135 "goal" attributes (40), we only used five sentences). We generated these sentences in a way that covered the diverse types of properties and features underpinning the attribute in consideration, as specified in prior published work (note that for the person perception sentences we simply used the ten associated traits specified in the article (45)). We also tried to diversify the sets of words and phrases used to maximize robustness.

We encoded all GPT-generated sentences as well as all attribute-related sentences as sentence embeddings (33, 59), and measured the degree to which a GPT-generated cost or benefit corresponded to the negative or positive instantiation of an attribute by calculating its average cosine similarity with the sentences describing that attribute in embedding space. Finally, we averaged each of the embedding similarities for each of the three costs or benefits, to get, for each choice option in our dataset, a 207-dimensional attribute representation of its costs and a 207-dimensional vector representation of its benefits. Each element in these vectors ranged from -1 to 1, with high values indicating that the GPT-generated sentences for the option were very similar in embedding space to the attribute associated with that dimension. See SOM 1.5.

**Validation Studies**

**Study 2.** In Study 2, we gave 50 participants (63% male, mean age = 27) recruited from Prolific Academic five randomly selected dilemmas from r/Advice, as well as GPT's outputs for each of the dilemmas. Participation in this study was restricted to individuals whose primary language was English, who had an approval rate on Prolific Academic greater than 98% and who had participated in 20 or more prior studies on the platform. Note that one participant timed out, resulting in a final sample size of 49 (this minor attrition should not affect our results). There were 13 GPT outputs in total (all textual). The first output corresponded to GPT's description of the two choice options in the dilemma, and the remaining twelve corresponded to GPT's description of the costs and benefits of the two options in the dilemma. Each participant was asked to provide a binary response indicating whether or not each of GPT's outputs was accurate or inaccurate. Since Study 1 used 100 total dilemmas, we obtained an average of 2.45 judgments for each GPT output. We calculated the modal participant judgment for each output (i.e. whether or not the participants, on average, judged the output to be accurate or inaccurate), and used this modal judgment in the analysis shown in Figure 2D. Outputs that were judged by an equal number of participants to be accurate vs. inaccurate were excluded from this analysis. Details of experimental materials are in SOM 3.2.

**Study 3.** In Study 3, we gave 100 participants (63% male, mean age = 34) recruited from Prolific Academic (with similar restrictions to Study 2) a set of 20 randomly selected attributes, as well as pairs of GPT-generated reasons that were coded by our sentence embeddings model to be either high or low on that attribute (there was no attrition in this study). Participants were asked to judge which of the GPT-generated reasons described something that had the target attribute. We generated stimuli for this study by taking, for each of the 207 attributes, 20 GPT-generated

reasons from our dataset that were the most similar and 20 that were the least similar to the attribute, as assessed by cosine similarity. Then, for each of the 207 attributes, we randomly picked three reasons from the high similarity group and three from the low similarity group, and paired them. This gave us a total of 621 unique judgment problems. We obtained an average of 3.22 participants judgments for each of these problems, resulting in an average of 9.66 judgments per attribute. We averaged these judgments to calculate the modal judgment for each attribute. These modal judgments were again averaged to generate the results shown in Figure 2D. These results present the proportion of modal judgments for attributes that identify the high-similarity reason (as assessed by our sentence embeddings) to be the most similar to the target attribute. Details of experimental materials are in SOM 3.3.

**Choice Prediction Studies**

**Studies 4a and b.** In Study 4a we asked 300 participants from Prolific Academic (51% male, mean age = 31) to make choices for eight r/Advice dilemmas (there was no attrition in this study). Participation in this study was restricted to individuals whose primary language was English, who has an approval rate on Prolific Academic greater than 99% and who had participated in 15 or more prior studies on the platform. The eight dilemmas used in this study were those that involved the highest tradeoff between the family closeness and security (FCS) cluster and pragmatism and financial prudence (PFP) cluster. We chose this pair of clusters since it was the most common tradeoff in our r/Advice dataset, excluding the sex and romance cluster (which we avoided due to the often explicit nature of its dilemmas). For each of the eight dilemmas, each participant was asked to imagine that they were facing the dilemma, choose between the two options in the dilemma (or indicate that they cannot choose) and also indicate their preference for the options on a 7-point Likert scale (ranging from strongly preferring the first option to strongly preferring the second). The study procedure was preregistered at https://osf.io/9ebrj, and although we also preregistered our analysis approach, here we present a different (but related) set of model-based analyses, in response to reviewer suggestions.

In Study 4b we asked 300 participants from Prolific Academic (53% male, mean age = 32) to make choices in eight additional r/Advice dilemmas (there was no attrition in this study). These were dilemmas that had the highest tradeoffs on the marital fulfilment vs. money and finance attribute clusters. We chose these since, excluding the clusters in Study 4a (and the sex and romance cluster) this was the most frequent tradeoff in r/Advice. We selected participants, constructed stimuli, and elicited responses in Study 4b in a manner that was identical to Study 4a (except that participation in Study 4b was restricted to individuals who had not taken Study 4a). This study was preregistered at https://osf.io/9ebrj.

Details of stimuli generation are in SOM 1.4, and details of models tested are in SOM 2.1. to 2.4. Details of experimental materials are in SOM 3.4.

**Studies 5a and b.** In Study 5a we recruited 302 participants from Prolific Academic (45% male, mean age = 30), with the same participation criteria in Studies 4a and b (there was no attrition in this study). Participants were asked to make a single choice involving the dilemma from Study 4a with the closest to 50% choice rate for option 1 vs 2. Each participant was asked to imagine that they were facing the dilemma, list the thoughts that come to mind as they deliberate in the order in which they come to mind, choose between the two options in the dilemma (or indicate that they cannot choose), and indicate their preference for the options on a 7-point Likert scale (ranging from strongly preferring the first option to strongly preferring the second). After this, participants

were shown each of their listed thoughts and were asked to indicate, for each option, whether the thought indicated a benefit for that option, a cost for that option, or neither. The order of the options was counterbalanced. This study was preregistered at https://osf.io/98gzm. In Study 5b we replicated Study 5a with the same recruitment criteria and procedure as Study 5a, except that we picked the dilemma from Study 4b and restricted participation to individuals who had not taken part in Study 5a. We recruited 302 participants from Prolific Academic (48% male, mean age = 32) (there was no attrition in this study). This study was also preregistered at https://osf.io/98gzm. Details of models tested are in SOM 2.5. to 2.7. Details of experimental materials are in SOM 3.5.

**References**


1. J. T. Austin, J. B. Vancouver, Goal constructs in psychology: Structure, process, and content. *Psychol. Bull.* **120**, 338–375 (1996).
2. J. R. Bettman, E. J. Johnson, J. W. Payne, “Consumer Decision Making” in *Handbook of Consumer Behavior*, (Prentice-Hall, Inc, 1991).
3. S. Bhatia, G. Loomes, D. Read, Establishing the laws of preferential choice behavior. *Judgm. Decis. Mak.* **16**, 1324–1369 (2021).
4. C. F. Camerer, G. Loewenstein, M. Rabin, *Advances in Behavioral Economics* (Princeton University Press, 2004).
5. E. Fehr, H. Gintis, Human motivation and social cooperation: Experimental and analytical foundations. *Annu. Rev. Sociol.* **33**, 43–64 (2007).
6. G. Gigerenzer, W. Gaissmaier, Heuristic Decision Making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
7. P. W. Glimcher, *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics* (MIT Press, 2004).
8. J. J. Bavel, *et al.*, Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).
9. S. Mertens, M. Herberz, U. J. J. Hahnel, T. Brosch, The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proc. Natl. Acad. Sci.* **119**, e2107346118 (2022).
10. K. L. Milkman, *et al.*, Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
11. R. H. Thaler, C. R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness* (Penguin, 2009).
12. L. R. Anderson, J. M. Mellor, Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. *J. Risk Uncertain.* **39**, 137–160 (2009).

13. R. Frey, A. Pedroni, R. Mata, J. Rieskamp, R. Hertwig, Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* **3**, e1701381 (2017).
14. A. Pedroni, *et al.*, The risk elicitation puzzle. *Nat. Hum. Behav.* **1**, 803–809 (2017).
15. M. Amlung, L. Vedelago, J. Acker, I. Balodis, J. MacKillop, Steep delay discounting and addictive behavior: a meta-analysis of continuous associations. *Addiction* **112**, 51–62 (2017).
16. M. Amlung, *et al.*, Delay Discounting as a Transdiagnostic Process in Psychiatric Disorders: A Meta-analysis. *JAMA Psychiatry* **76**, 1176–1186 (2019).
17. A. J. Bailey, R. J. Romeu, P. R. Finn, The problems with delay discounting: a critical review of current practices and clinical applications. *Psychol. Med.* **51**, 1799–1806 (2021).
18. D. M. Bartels, Y. Li, S. Bharti, How well do laboratory-derived estimates of time preference predict real-world behaviors? Comparisons to four benchmarks. *J. Exp. Psychol. Gen.* **152**, 2651–2665 (2023).
19. S. T. Van Baal, J. Hohwy, A. Verdejo-García, E. Konstantinidis, L. Walasek, Fenneman et al.'s (2022) review of formal impulsivity models: Implications for theory and measures of impulsivity. *Psychol. Bull.* (2023). https://doi.org/10.1037/bul0000404.
20. B. Szaszi, *et al.*, No reason to expect large and consistent effects of nudge interventions. *Proc. Natl. Acad. Sci.* **119**, e2200732119 (2022).
21. S. DellaVigna, E. Linos, RCTs to Scale: Comprehensive Evidence From Two Nudge Units. *Econometrica* **90**, 81–116 (2022).
22. M. Muthukrishna, J. Henrich, A problem in theory. *Nat. Hum. Behav.* **3**, 221–229 (2019).
23. R. M. Shiffrin, Is it Reasonable to Study Decision-Making Quantitatively? *Top. Cogn. Sci.* **14**, 621–633 (2022).
24. S. Vazire, S. R. Schiavone, J. G. Bottesini, Credibility beyond replicability: Improving the four validities in psychological science. *Curr. Dir. Psychol. Sci.* **31**, 162–168 (2022).

25. L. Walasek, G. D. A. Brown, Incomparability and Incommensurability in Choice: No Common Currency of Value? *Perspect. Psychol. Sci.* 17456916231192828 (2023). https://doi.org/10.1177/17456916231192828.
26. N. Chater, G. Loewenstein, The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behav. Brain Sci.* **46**, e147 (2023).
27. S. Bhatia, M. Galesic, M. Mitchell, Editorial for the Special Issue on Algorithms in Our Lives. *Perspect. Psychol. Sci.* 17456916231214452 (2024). https://doi.org/10.1177/17456916231214452.
28. T. L. Griffiths, Manifesto for a new (computational) cognitive revolution. *Cognition* **135**, 21–23 (2015).
29. S. C. Matz, O. Netzer, Using Big Data as a window into consumers' psychology. *Curr. Opin. Behav. Sci.* **18**, 7–12 (2017).
30. H. A. Schwartz, *et al.*, Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* **8**, e73791 (2013).
31. S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, J. C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* **18**, 43–49 (2017).
32. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, T. Solorio, Eds. (Association for Computational Linguistics, 2019), pp. 4171–4186.
33. K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: Masked and Permuted Pre-training for Language Understanding in *Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2020), pp. 16857–16867.

34. T. Brown, *et al.*, Language Models are Few-Shot Learners in *Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2020), pp. 1877–1901.

35. D. Demszky, *et al.*, Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).

36. M. Deghani, R. L. Boyd, *Handbook of Language Analysis in Psychology* (Guildford Press, 2022).

37. S. Bhatia, A. Aka, Cognitive Modeling With Representations From Large-Scale Digital Data. *Curr. Dir. Psychol. Sci.* **31**, 207–214 (2022).

38. I. Grossmann, *et al.*, AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).

39. S. Bhatia, R. Richie, W. Zou, Distributed semantic representations for modeling human judgment. *Curr. Opin. Behav. Sci.* **29**, 31–36 (2019).

40. A. S. Chulef, S. J. Read, D. A. Walsh, A Hierarchical Taxonomy of Human Goals. *Motiv. Emot.* **25**, 191–232 (2001).

41. P. A. M. Van Lange, The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *J. Pers. Soc. Psychol.* **77**, 337–349 (1999).

42. K. E. Voss, E. R. Spangenberg, B. Grohmann, Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude. *J. Mark. Res.* **40**, 310–320 (2003).

43. J. A. Russell, A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).

44. S. Schwartz, W. Bilsky, Toward A Universal Psychological Structure of Human Values. *J. Pers. Soc. Psychol.* **53**, 550–562 (1987).

45. G. P. Goodwin, J. Piazza, P. Rozin, Moral character predominates in person perception and evaluation. *J. Pers. Soc. Psychol.* **106**, 148–168 (2014).

46. B. Harward, A. Taylor, Kavanagh, Shane, What's Fair? The Three Forms of Fairness. (2021). Available at: https://www.gfoa.org/materials/whats-fair-1 [Accessed 26 January 2024].

47. R. M. Ryan, E. L. Deci, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **55**, 68–78 (2000).
48. J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**, 1029–1046 (2009).
49. M. W. Krawczyk, A model of procedural and distributive fairness. *Theory Decis.* **70**, 111–128 (2011).
50. H. Karnofsky, Update on Cause Prioritization at Open Philanthropy. *Open Philanthr.* (2018). Available at: https://www.openphilanthropy.org/research/update-on-cause-prioritization-at-open-philanthropy/ [Accessed 4 February 2024].
51. P. Singer, *10th Anniversary Edition The Life You Can Save: How To Do Your Part To End World Poverty* (The Life You Can Save.org, 2019).
52. T. Gilovich, I. Gallo, Consumers' pursuit of material and experiential purchases: A review. *Consum. Psychol. Rev.* **3**, 20–33 (2020).
53. P. Ekman, "Basic emotions" in *Handbook of Cognition and Emotion*, (John Wiley & Sons Ltd, 1999), pp. 45–60.
54. S. T. Fiske, A. J. C. Cuddy, P. Glick, Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
55. R. L. Keeney, H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Trade-Offs* (Cambridge University Press, 1993).
56. R. M. Hogarth, N. Karelaia, Simple Models for Multiattribute Choice with Many Alternatives: When It Does and Does Not Pay to Face Trade-offs with Binary Attributes. *Manag. Sci.* **51**, 1860–1872 (2005).
57. S. Bhatia, N. Stewart, Naturalistic multiattribute choice. *Cognition* **179**, 71–88 (2018).
58. J. R. Busemeyer, S. Gluth, J. Rieskamp, B. M. Turner, Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends Cogn. Sci.* **23**, 251–263 (2019).

59. N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, X. Wan, Eds. (Association for Computational Linguistics, 2019), pp. 3982–3992.
60. E. U. Weber, *et al.*, Asymmetric Discounting in Intertemporal Choice: A Query-Theory Account. *Psychol. Sci.* **18**, 516–523 (2007).
61. M. Schulte-Mecklenbeck, A. Kuehberger, R. Raynard, Eds., *A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide* (Psychology Press, 2010).
62. K. A. Ericsson, H. A. Simon, Verbal reports as data. *Psychol. Rev.* **87**, 215–251 (1980).
63. B. Rammstedt, O. P. John, Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J. Res. Personal.* **41**, 203–212 (2007).
64. A. R. Camilleri, An investigation of big life decisions. *Judgm. Decis. Mak.* **18**, e32 (2023).
65. L. A. Paul, *Transformative Experience* (OUP Oxford, 2014).
66. S. Hechtlinger, C. Schulze, C. Leuker, R. Hertwig, The psychology of life's most important decisions. *Am. Psychol.* (2024). https://doi.org/10.1037/amp0001439.
67. Q. J. M. Huys, *et al.*, Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLOS Comput. Biol.* **8**, e1002410 (2012).
68. I. Krajbich, A. Rangel, Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc. Natl. Acad. Sci.* **108**, 13852–13857 (2011).
69. T. Noguchi, N. Stewart, Multialternative decision by sampling: A model of decision making constrained by process data. *Psychol. Rev.* **125**, 512–544 (2018).
70. E. Goffman, *The presentation of self in everyday life* (Doubleday, 1959).

71. R. F. Baumeister, A self-presentational view of social phenomena. *Psychol. Bull.* **91**, 3–26 (1982).

72. B. R. Schlenker, *Impression management: the self-concept social identity, and interpersonal relations* (Brooks/Cole, 1980).

73. P. Johansson, L. Hall, S. Sikström, A. Olsson, Failure to detect mismatches between intention and outcome in a simple decision task. *Science* **310**, 116–119 (2005).

74. T. Wilson, *Strangers to Ourselves* (2004).

75. A. G. Greenwald, D. E. McGhee, J. L. Schwartz, Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).

76. S. Bhatia, L. Walasek, Predicting implicit attitudes with natural language data. *PNAS Proc. Natl. Acad. Sci. U. S. Am.* **120**, 1–8 (2023).

77. T. D. Wilson, The Proper Protocol: Validity and Completeness of Verbal Reports. *Psychol. Sci.* **5**, 249–252 (1994).

78. E. U. Weber, E. J. Johnson, Query theory: Knowing what we want by arguing with ourselves. *Behav. Brain Sci.* **34**, 91–92 (2011).

79. F. Wang, A. Aka, L. He, S. Bhatia, Memory modeling of counterfactual generation. *J. Exp. Psychol. Learn. Mem. Cogn.* https://doi.org/10.1037/xlm0001335.

80. A. Morris, R. W. Carlson, H. Kober, M. Crockett, Introspective access to value-based multi-attribute choice processes. [Preprint] (2023). Available at: https://osf.io/2zrfa [Accessed 18 November 2024].