Indecision Modeling

Duncan C McElfresh,[1] Lok Chan,[2] Kenzie Doyle,[3]
Walter Sinnott-Armstrong,[2] Vincent Conitzer,[2] Jana Schaich Borg,[2] John P Dickerson[1]
[1] University of Maryland, College Park
[2] Duke University
[3] University of Oregon
dmcelfre@umd.edu, lok.c@duke.edu

Abstract

AI systems are often used to make or contribute to important decisions in a growing range of applications, including criminal justice, hiring, and medicine. Since these decisions impact human lives, it is important that the AI systems act in ways which align with human values. Techniques for preference modeling and social choice help researchers learn and aggregate peoples' preferences, which are used to guide AI behavior; thus, it is imperative that these learned preferences are accurate. These techniques often assume that people are willing to express strict preferences over alternatives; which is not true in practice. People are often indecisive, and especially so when their decision has moral implications. The philosophy and psychology literature shows that indecision is a measurable and nuanced behavior—and that there are several different reasons people are indecisive. This complicates the task of both learning and aggregating preferences, since most of the relevant literature makes restrictive assumptions on the meaning of indecision. We begin to close this gap by formalizing several mathematical *indecision* models based on theories from philosophy, psychology, and economics; these models can be used to describe (indecisive) agent decisions, both when they are allowed to express indecision and when they are not. We test these models using data collected from an online survey where participants choose how to (hypothetically) allocate organs to patients waiting for a transplant.

Introduction

AI systems are currently used to make, or contribute to, many important decisions. These systems are deployed in self-driving cars, organ allocation programs, businesses for hiring, and courtrooms to set bail. It is an ongoing challenge for AI researchers to ensure that these systems make decisions that align with human values.

A growing body of research views this challenge through the lens of *preference aggregation*. From this perspective, researchers aim to (1) understand the preferences (or values) of the relevant *stakeholders*, and (2) design an AI system that aligns with the aggregated preferences of all stakeholders. This approach has been proposed recently in the context of self-driving cars (Noothigattu et al. 2018) and organ allocation (Freedman et al. 2020). These approaches rely on a mathematical model of stakeholder preferences–which is typically *learned* using data collected via hypothetical decision scenarios or online surveys.[1] There is a rich literature addressing how to elicit preferences accurately and efficiently, spanning the fields of computer science, operations research, and social science.

It is critical that these *observed* preferences accurately represent peoples' *true* preferences, since these observations guide deployed AI systems. Importantly, the way we measure (or *elicit*) preferences is closely tied to the accuracy of these observations. In particular, it is well-known that both the order in which questions are asked, and the set of choices presented, impact expressed preferences (Day et al. 2012; DeShazo and Fermo 2002).

Often people choose *not* to express a strict preference, in which case we call them *indecisive*. The economics literature has suggested a variety of explanations for indecision (Gerasimou 2018)—for example when there are no desirable alternatives, or when all alternatives are perceived as equivalent. Moral psychology research has found that people often "do not want to play god" in moral situations, and would prefer for somebody or something else to take responsibility for the decision (Gangemi and Mancini 2013).

In philosophy, indecision of the kind discussed in this paper is typically linked to a class of moral problems called symmetrical dilemmas, in which an agent is confronted with the choice between two alternatives that are or appear to the agent equal in value (Sinnott-Armstrong 1988).[2] Much of the literature concerns itself with the morality and rationality of the use of a randomizer, such as flipping a coin, to resolve these dilemmas. Despite some disagreements over details (McIntyre 1990; Donagan 1984; Blackburn 1996; Hare 1981), many philosophers do agree that flipping a coin is often a viable course of action in response to indecision[3].

The present study accepts the assumption that flipping a coin is typically an expression of one's preference to *not* decide between two options, but goes beyond the received

[1] The MIT Moral Machine project is one example: https://www.moralmachine.net/

[2] Sophie's Choice is a well-known example: a guard at the concentration camp cruelly forces Sophie to choose one of her two children to be killed. The guard will kill both children if Sophie refuses to choose. Sophie's reason for not choosing one child applies equally to another, hence the symmetry.

[3] With some exceptions: for example, see (Railton 1992).

view in philosophy by suggesting that indecision can also be common and acceptable when the alternatives are *asymmetric*. We show that people often do adopt coin flipping strategies in asymmetrical dilemmas, where the alternatives are not equal in value. Thus, the use of a randomizer is likely to play a more complex role in moral decision making than simply as a tie breaker for symmetrical dilemmas.

Naturally, people are also sometimes indecisive when faced with difficult decisions related to AI systems. However it is commonly assumed in the preference modeling literature that people always express a strict preference, unless (A) the alternatives are approximately equivalent, or (B) the alternatives are incomparable. Assumption (A) is mathematically convenient, since it is necessary for preference *transitivity*.[4] Since indecision is both a common and meaningful response, strict preferences alone cannot accurately represent peoples' real values. Thus, AI researchers who wish to guide their systems using observed preferences should be aware of the hidden meanings of indecision. We aim to uncover these meanings in a series of studies.

Our Contributions. First, we conduct a pilot experiment to illustrate how different interpretations of indecision lead to different outcomes (§). Using hypothesis testing, we reject the common assumption (A) that indecision is expressed only toward equivalent—or symmetric—alternatives.

Then, drawing on ideas from psychology, philosophy, and economics, we discuss several other potential reasons for indecision, drawing (§). We formalize these ideas as mathematical indecision *models*, and develop a probabilistic interpretation that lends itself to computation (§).

To test the utility of these models, we conduct a second experiment to collect a much larger dataset of decision responses (§). We take a machine learning (ML) perspective, and evaluate each model class based on its goodness-of-fit to this dataset. We assess each model class for predicting *individual* peoples' responses, and then we briefly investigate group decision models.

In all of our studies, we ask participants *who should receive the kidney?* in a hypothetical scenario where two patients are in need of a kidney, but only one kidney is available. As a potential basis for their answers, participants are given three "features" of each patient: age, amount of alcohol consumption, and number of young dependents.

We chose this task for several reasons: first, kidney exchange is a real application where algorithms influence—and sometimes make—important decisions about who receives which organ.[5] Second, organ allocation is a *difficult* problem: there are far fewer donors organs than there are people in need of a transplant.[6] Third, the question of *who* should receive these scarce resources raises serious ethical dilemmas (Scheunemann and White 2011). Kidney allocation is also a common motivation for studies of fair resource allocation (Agarwal et al. 2019; McElfresh and Dickerson 2018; Mattei, Saffidine, and Walsh 2018). Furthermore, this type of scenario is frequently used to study peoples' preferences and behavior (Freedman et al. 2020; Furnham, Simmons, and McClelland 2000; Furnham, Thomson, and McClelland 2002; Oedingen, Bartling, and Krauth 2018). Importantly, this prior work focuses on peoples' *strict* preferences, while we aim to study indecision.

[4]My preferences are transitive if "I prefer A over B" and "I prefer B over C" implies "I prefer A over C".

[5]Many exchanges match patients and donors algorithmically, including the United Network for Organ Sharing (https://unos.org/transplant/kidney-paired-donation/) and the UK national exchange (https://www.odt.nhs.uk/living-donation/uk-living-kidney-sharing-scheme/).

[6]There are around $100,000$ people in need of a transplant today (https://unos.org/data/transplant-trends/), and about $22,000$ transplants have been conducted in 2020.

Study 1: Indecision is Not Random Choice

We first conduct a pilot study to illustrate the importance of measuring indecision. Here we take the perspective of a preference-aggregator; we illustrate this perspective using a brief example: Suppose we must choose between two alternatives (X or Y), based on the preferences of several stakeholders. Using a survey we ask all stakeholders to express a *strict* preference (to "vote") for their preferred alternative; X receives 10 votes while Y receives 6 votes, so X wins.

Next we conduct the same survey, but allow stakeholders to vote for "indecision" instead; now, X receives 4 votes, Y receives 5 votes, and "indecision" receives 7 votes. If we assume that voters are indecisive only when alternatives are nearly equivalent (assumption (A) from Section), then each "indecision" vote is analogous to one half-vote for both X and Y, and therefore Y wins. In other words, in the first survey we assume that all indecisive voters choose randomly between X and Y. However, if indecision has another meaning, then it is not clear whether X or Y wins. Thus, in order to make the best decision for our constituents we must understand what meaning is conveyed by indecisive voters. Unfortunately for our hypothetical decision-maker, assumption (A) is not always valid.

Using a small study, we test—and reject—assumption (A), which we frame as two different hypotheses, **H0-1**: *if we discard all indecisive votes, then both X and Y receive the same proportion votes, whether or not indecision is allowed.* A second related hypothesis is **H0-2**: *if we assign half of a vote to both X and Y when someone is indecisive, then both X and Y receive the same proportion votes, whether or not indecision is allowed.* We conducted the hypothetical surveys described above, using 15 kidney allocation questions (see Appendix **??** for the survey text and analysis). Participants were divided into two groups: participants in group *Indecisive* (N=62) were allowed to express indecision (phrased as "flip a coin to decide who receives the kidney"), while group *Strict* (N=60) was forced to choose one of the two recipients. We test **H0-1** by identifying the *majority patient*, "X" (who received the most votes) and the *minority patient* "Y" for each of the 15 questions (details of this analysis are in Appendix **??**). Overall, group *Indecisive* cast 581 (74) votes for the majority (minority) patient, and 275 indecision votes; the *Strict* group cast 751 (149) votes for the majority (minority) patient. Using a Pearson's chi-squared test we reject **H0-1** ($p < 0.01$). According to **H0-2**, we might assume that all indecision votes are "effectively" one half-vote for both

the minority and majority patient. In this case, the *Indecisive* group casts 718.5 (211.5) "effective" votes for the majority (minority) patients; using these votes we reject **H0-2** ($p < 0.01$).

In the context of our hypothetical choice between X and Y, this finding is troublesome: since we reject **H0-1** and **H0-2**, we cannot choose a winner by selecting the alternative with the most votes—or, if indecision is measured, the most "effective" votes. If indecision has other meanings, then the "best" alternative depends on which meanings are used by each person; this is our focus in the remainder of this paper.

Models for Indecision

The psychology and philosophy literature find several reasons for indecision, and many of these reasons can be approximated by numerical decision models. Before presenting these models, we briefly discuss their related theories from psychology and philosophy.

Difference-Based Indecision In the preference modeling literature it is sometimes assumed that people are indecisive only when both alternatives (X and Y) are indistinguishable. That is, the perceived difference between X and Y is too small to arrive at a strict preference. In philosophy, this is referred to as "the possibility of parity" (Chang 2002).

Desirability-Based Indecision In cases where both alternatives are not "good enough", people may be reluctant to choose one over the other. This has been referred to as "single option aversion" (Mochon 2013), when consumers do not choose between product options if none of the options is sufficiently likable. Zakay (1984) observes this effect in single-alternative choices: people reject an alternative if it is not sufficiently close to a hypothetical "ideal". Similarly, people may be indecisive if *both* alternatives are attractive. People faced with the choice between two highly valued options often opt for an indecisive resolution in order to manage negative emotions (Luce 1998).

Conflict-Based Indecision People may be indecisive when there are both good and bad attributes of each alternative. This is phrased as *conflict* by Tversky and Shafir (1992): people have trouble deciding between two alternatives if neither is better than the other in *every way*. In the AI literature, the concept of *incomparability* between alternatives is also studied (Pini et al. 2011).

While these notions are intuitively plausible, we need mathematical definitions in order to model observed preferences. That is the purpose of the next section.

Indecision Model Formalism

In accordance with the literature, we refer to decision-makers as *agents*. Agent preferences are represented by binary relations over each pair of items $(i, j) \in \mathcal{I} \times \mathcal{I}$, where $\mathcal{I}$ is a universe of items. We assume agent preferences are *complete*: when presented with item pair (i, j), they expresses exactly one response $r \in \{0, 1, 2\}$, which indicates:

- $r = 1$, or $i \succ j$: the agent prefers i more than j
- $r = 2$, or $i \prec j$: the agent prefers j more than i
- $r = 0$, or $i \sim j$: the agent is indecisive between i and j

When preferences are complete and transitive,[7] then the preference relation corresponds to a weak ordering over all items (Shapley and Shubik 1974). In this case there is a utility function representation for agent preferences, such that $i \succ j \iff u(i) > u(j)$, and $i \sim j \iff u(i) = u(j)$, where $u : \mathcal{I} \to \mathbb{R}$ is a continuous function. We assume each agent has an underlying utility function, however in general we *do not* assume preferences are transitive. In other words, we assume agents can rank items based on their relative value (represented by $u(\cdot)$), but in some cases they consider other factors in their response—causing them to be indecisive. Next, to model indecision we propose mathematical representations of the causes for indecision from Section .

Mathematical Indecision Models

All models in this section are specified by two parameters: a utility function $u(\cdot)$ and a threshold λ. Each model is based on *scoring functions*: when the agent observes a query they assign a numerical score to each response, and they respond with the response type that has maximal score; we assume that score ties are broken randomly, though this assumption will not be important. In accordance with the literature, we assume the agent observes random iid additive error for each response score (see, e.g., Soufiani, Parkes, and Xia (2013)). Let $S_r(i, j)$ be the agent's score for response r to comparison (i, j); the agent's response is given by

$$R(i, j) = \operatorname*{argmax}_{r \in \{0,1,2\}} S_r(i, j) + \epsilon_{rij}.$$

That is, the agent has a deterministic score for each response $S_r(i, j)$, but when making a decision the agent observes a noisy version of this score, $S_r(i, j) + \epsilon_{rij}$. We make the common assumption that noise terms ϵ_{rij} are iid Gumbel-distributed, with scale $\mu = 1$. In this case, the distribution of agent responses is

$$p(i, j, r) = \frac{e^{S_r(i,j)}}{e^{S_0(i,j)} + e^{S_1(i,j)} + e^{S_2(i,j)}}. \tag{1}$$

Each indecision model is defined using different score functions $S_r(\cdot, \cdot)$. Score functions for strict responses are always symmetric, in the sense that $S_2(i, j) = S_1(j, i)$; thus we need only define $S_1(\cdot, \cdot)$ and $S_0(\cdot, \cdot)$. We group each model by their cause for indecision from Section .

Difference-Based Models: `Min-`δ, `Max-`δ Agents are indecisive when the utility difference between alternatives is either smaller than threshold λ (`Min-`δ) or greater than λ (`Max-`δ). The score functions for these models are

$$\texttt{Min-}\delta: \begin{cases} S_1(i, j) & \equiv u(i) - u(j) \\ S_0(i, j) & \equiv \lambda \end{cases}$$

$$\texttt{Max-}\delta: \begin{cases} S_1(i, j) & \equiv u(i) - u(j) \\ S_0(i, j) & \equiv 2|u(i) - u(j)| - \lambda \end{cases}$$

Here λ should be non-negative: for example with `Min-`δ, $\lambda \leq 0$ means the agent is never indecisive, while for `Max-`δ this means the agent is always indecisive. Model `Max-`δ seems counter-intuitive (if one alternative is clearly better than the other, why be indecisive?), yet we include it

[7] Agent preferences are transitive if $i \succ j$ and $i \succ k$ iff $i \succ k$.

for completeness. Note that this is only one example of a difference-based model: instead the agent might assess alternatives using a distance measure $d : \mathcal{I} \times \mathcal{I} \to \mathbb{R}_+$, rather than $u(\cdot)$.

Desirability-Based Models: `Min`$-U$, `Max`$-U$ Agents are indecisive when the utility of *both* alternatives is below threshold λ (`Min`$-U$), or when the utility of both alternatives is greater than λ (`Max`$-U$). Unlike the difference-based models, λ here may be positive or negative. The score functions for these models are

$$\texttt{Min}\text{-}U: \begin{cases} S_1(i,j) & \equiv u(i) \\ S_0(i,j) & \equiv \lambda \end{cases}$$

$$\texttt{Max}\text{-}U: \begin{cases} S_1(i,j) & \equiv u(i) \\ S_0(i,j) & \equiv 2\min\{u(i),u(j)\} - \lambda \end{cases}$$

Both of these models motivated in the literature (see §).

Conflict-Based Model: `Dom` In this model the agent is indecisive unless one alternative *dominates* the other in all features, by threshold at least λ. For this indecision model, we need a utility measure associated with each feature of each item; for this purpose, let $u_n(i)$ be the utility associated with feature n of item i. As before, λ here may be positive or negative. The score functions for this model are

$$\texttt{Dom}: \begin{cases} S_1(i,j) & \equiv \min_{n\in[N]} \left(u_n(i) - u_n(j)\right) \\ S_0(i,j) & \equiv \lambda \end{cases}$$

This is one example of a conflict-based indecision model, though we might imagine others.

These models serve as a class of hypotheses which describe how agents respond to comparisons when they are allowed to be indecisive. Using the response distribution in (1), we can assess how well each model fits with an agent's (possibly indecisive) responses. However, in many cases agents are *required* to express strict preferences—they are not allowed to be indecisive (as in Section). With slight modification the score-based models from this section can be used even when agents are forced to express *only* strict preferences; we discuss this in the next section.

Indecision Models for Strict Comparisons

We assume that agents may prefer to be indecisive, even when they are required to express strict preferences. That is, we assume that agents use an underlying *indecision* model to express *strict* preferences. When they cannot express indecision, we assume that they either *resample* from their decision distribution, or they choose randomly. That is, we assume agents use a two-stage process to respond to queries: first they sample a response r from their response distribution $p(\cdot,\cdot,r)$; if r is strict (1 or 2), then they express it, and we are done. If they sample indecision (0), then they flip a weighted coin to decide how to respond:

(heads) with probability q they re-sample from their response distribution until they sample a strict response, without flipping the weighted coin again

(tails) with probability $1-q$ they choose uniformly at randomly between responses 1 and 2.

That is, they respond according to distribution

$$p_{strict}(i,j,r) \equiv \begin{cases} q\left(\frac{e^{S(i,j)}+(1/2)e^{S_0(i,j)}}{C}\right) & \\ \quad +\frac{1-q}{D}\left(e^{S_1(i,j)}\right) & \text{if } r=1 \\ q\left(\frac{e^{S_2(i,j)}+(1/2)e^{S_0(i,j)}}{C}\right) & \\ \quad +\frac{1-q}{D}e^{S_2(i,j)} & \text{if } r=2 \end{cases} \tag{2}$$

Here, $C \equiv e^{S_0(i,j)} + e^{S_1(i,j)} + e^{S_2(i,j)}$, and $D \equiv e^{S_1(i,j)} + e^{S(_2(i,j)}$. The (heads) condition from above has another interpretation: the agent chooses to sample from a "strict" logit, induced by only the score functions for strict responses, $S_1(i,j)$ and $S_1(i,j)$. We discuss this model in more detail, and provide an intuitive example, in Appendix **??**.

We now have mathematical indecision models which describe how indecisive agents respond to comparison queries, both when they are allowed to express indecision (§), and when they are not (§). The model in this section, and response distributions (1) and (2), represent one way indecisive agents might respond when they are forced to express strict preferences. The question remains whether any of these models accurately represent peoples' expressed preferences in real decision scenarios. In the next section we conduct a second, larger survey to address this question.

Study 2: Fitting Indecision Models

In our second study, we aim to *model* peoples' responses in the hypothetical kidney allocation scenario using indecision models from the previous section as well as standard preference models from the literature. The models from the previous section can be used to predict peoples' responses, both when they are allowed to be indecisive, and when they are not. To test both class of models, we conducted a survey with two groups of participants, where one group was were given the option to express indecision, and the other was not. Each participant was assigned to 1 of the 150 random sequences, each of which contains 40 pairwise comparisons between two hypothetical kidney recipients with randomly generated values for age, number of dependents, and number of alcoholic drinks per week. We recruited 150 participants for group *Indecisive*, which was given the option to express indecision[8]. 18 participants were excluded from the analysis for failing attention checks, leaving us with a final sample of N=132. Another group, *Strict* (N=132), was recruited to respond to the same 132 sequences, but without the option to express indecision.

We remove 26 participants from *Indecisive* who never express indecision, because it is not sensible to compare goodness-of-fit for different indecision models when the agent never chooses to be indecisive. This study was reviewed and approved by our organization's Institutional Review Board; please see Appendix **??** for a full description of the survey and dataset.

Model Fitting. In order to fit these indecision models to data, we assume that agent utility functions are linear: each item $i \in \mathcal{I}$ is represented by feature vector $\boldsymbol{x}^i \in \mathbb{R}^N$;

[8] As in Study 1, this is phrased as "flip a coin."

agent utility for item i is $u(i) = \boldsymbol{u}^\top \boldsymbol{x}^i$, where $\boldsymbol{u} \in \mathbb{R}^N$ is the agent's *utility vector*. We take a maximum likelihood estimation (MLE) approach to fitting each model: i.e., we select agent parameters $\boldsymbol{u}$ and λ which maximize the log-likelihood (LL) of the training responses. Since the LL of these models is not convex, we use random search via a Sobol process (Sobol' 1967). The search domain for utility vectors is $\boldsymbol{u} \in [-1, 1]^N$, the domain for probability parameters is $(0, 1)$, and the domain for λ depends on the model type (see Appendix **??**). The number of candidate parameters tested and the nature of the train-test split vary between experiments. All code used for our analysis is available online, [9] and details of our implementation can be found in Appendix **??**.

We explore two different preference-modeling settings: learning individual indecision models, and learning group indecision models.

Individual Indecision Models

The indecision models from Section are indented to describe how an indecisive agent responds to queries—both when they are given the option to be indecisive, and when they are not. Thus, we fit each of these models to responses from both participant groups: *Indecisive* and *Strict*. For each participant we randomly split their question-response pairs into a training and testing set of equal size (20 responses each). For each participant we fit all five models from Section , and two baseline methods: `Rand` (express indecision with probability q and chooses randomly between alternatives otherwise), `MLP` (a multilayer perceptron classifier with two hidden layers with 32 and 16 nodes). We use `MLP` as a state-of-the-art benchmark, against which we compare our models; we use this benchmark to see how close our new models are to modern ML methods.

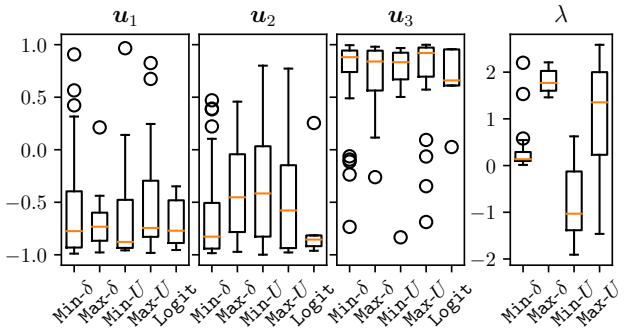
For group *Indecisive* we estimate parameter q for `NaiveRand` from the training queries; for *Strict* q is 0. For `MLP` we train a classifier with one class for each response type, using scikit-learn (Pedregosa et al. 2011): for *Indecisive* responses we train a three-class model ($r \in \{0, 1, 2\}$), and for *Strict* we train a two-class model ($r \in \{1, 2\}$).

Goodness-of-fit. Using the standard ML approach, we select the best-fit models for each agent using the training-set LL, and evaluate the performance of these best-fit models using the test-set LL. Table 1 shows the number of participants for which each model was the 1st-, 2nd-, and 3rd best-fit for each participant (those with the greatest training-set LL), and the median test and train LL for each model. First we observe that *no indecision model* is a clear winner: several different models appear in the top 3 for each participant. This suggests that different indecision models fit different individuals better than others — there is not a single model that reflects everyone's choices. However, some models perform better than others: `Min-`δ and `Max-`δ appear often in the top 3 models, as does `Max-`U for group *Indecisive*.

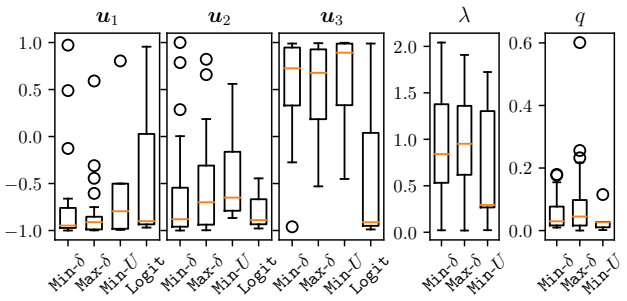
It it is somewhat surprising the `Max-`δ fits participant responses, since this model does not seem intuitive: in `Max-`δ, agents are indecisive when two alternatives have *very different* utility—i.e. one has much greater utility than the other. It is also surprising the `Max-`U is a good fit for group *Indecisive*, but not for *Strict*. One interpretation of this fact is that some people use (a version of) `Max-`U when they have the option, but they *do not* use `Max-`U when indecision is not an option. Another interpretation is that our modeling assumptions in Section are wrong—however our dataset cannot definitively explain this discrepancy.

Finally, `MLP` is the most common best-fit model for all participants in both groups, though it is rarely a 2nd- or 3rd-best fit. This suggests that the `MLP` benchmark accurately models *some* participants' responses, and performs poorly for others; we expect this is due to overfitting. While `MLP` is more accurate than our models in some cases, it does not shed light on why people are indecisive.

It is notable that some indecision models (`Min-`δ and `Max-`δ) outperform the standard logit model (`Logit`), both when they are learned from responses including indecision (group *Indecisive*), and when they are learned from only strict responses (group *Strict*). Thus, we believe that these indecision models give a more-accurate representation for peoples' decisions than the standard logit, both when they are given the option to be indecisive, and when they are not.



(a) Participant best-fit model parameters for *Indecisive*



(b) Participant best-fit model parameters for *Strict*

Figure 1: Best-fit parameters for each indecision model, for participants in group *Indecisive* (top) and *Strict* (bottom). Elements of the agent utility vector correspond to patient age ($\boldsymbol{u}_1$), alcohol consumption ($\boldsymbol{u}_2$), and number of dependents ($\boldsymbol{u}_3$); the interpretation of λ depends on the model class. Only participants for which the model is the 1st-best-fit are included (see Table 1).

[9] https://github.com/duncanmcelfresh/indecision-modeling

	Group *Indecisive* (both indecision and strict responses)				Group *Strict* (only strict responses)			
Model	#1st	#2nd	#3rd	Train/Test LL	# 1st	# 2nd	# 3rd	Train/Test LL
Min-δ	29 (27%)	23 (22%)	13 (12%)	-0.82/-0.85	26 (20%)	53 (40%)	34 (26%)	-0.44/-0.47
Max-δ	11 (10%)	12 (11%)	19 (18%)	-0.81/-0.90	31 (23%)	57 (43%)	25 (19%)	-0.44/-0.47
Min-U	8 (8%)	32 (30%)	17 (16%)	-0.83/-0.88	1 (1%)	5 (4%)	20 (15%)	-0.53/-0.56
Max-U	22 (21%)	23 (22%)	12 (11%)	-0.81/-0.83	1 (1%)	5 (4%)	15 (11%)	-0.53/-0.55
Dom	0 (0%)	3 (3%)	9 (8%)	-0.88/-0.95	2 (2%)	4 (3%)	3 (2%)	-0.57/-0.58
Logit	5 (5%)	12 (11%)	31 (29%)	-0.84/-0.90	4 (3%)	5 (4%)	27 (20%)	-0.53/-0.55
Rand	1 (1%)	0 (0%)	3 (3%)	-1.10/-1.10	6 (5%)	0 (0%)	1 (1%)	-0.69/-0.69
MLP	30 (28%)	1 (1%)	2 (2%)	-0.04/-1.15	61 (46%)	3 (2%)	7 (5%)	-0.03/-0.49

Table 1: Best-fit models for individual participants in group *Indecisive* (left) and *Strict* (right). The number of participants for which each model has the largest test log-likelihood (#1st), second-largest test LL (#2nd), as well as third-largest (#3rd) are given for each model, and the median training and test LL over all participants.

Since these indecision models may be accurate representations of peoples' choices, it is informative to examine the best-fit parameters. Figure 1 shows best-fit parameters for participants in group *Indecisive* (top) and *Strict* (bottom); for each indecision model, we show all learned parameters for participants for whom the model is the 1st-best-fit (see Table 1). Importantly, the best-fit values of $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ are similar for all models, in both groups. That is, *in general*, people have similar relative valuations for different alternatives: $\boldsymbol{u}_1 < 0$ means younger patients are preferred over older patients, $\boldsymbol{u}_2 < 0$ means patients who consume less alcohol are preferred more; $\boldsymbol{u}_3 > 0$ means that patients with more dependents are preferred more. We emphasize that the indecision model parameters for group *Strict* (bottom panel of Figure 1) are learned using only strict responses.

These models are fit using only 20 samples, yet they provide useful insight into how people make decisions. Importantly, our simple indecision models fit observed data better than the standard logit—both when people can express indecision, and when they cannot. Thus, contrary to the common assumption in the literature, not all people are indecisive *only* when two alternatives are nearly equivalent. This assumption may be true for some people (participants for which Min-δ is a best-fit model), but it is not always true.

Group Models

Next we turn to group decision models, where the goal is for an AI system to make decisions that reflect the values of a certain group of humans. In the spirit of the social choice literature, we refer to agents as "voters", and suggested decisions as "votes". We consider two distinct learning paradigms, where each reflects a potential use-case of an AI decision-making system.

The first paradigm, *Population Modeling*, concerns a large or infinite number of voters; our goal is to estimate responses to new decision problems that are the *best* for the entire population. This scenario is similar to conducting a national poll: we have a population including thousands or millions of voters, but we can only sample a small number (say, hundreds) of votes. Thus, we aim to build a model that represents the entire population, using a small number of votes from a small number of voters. There are several ways to aggregate uncertain voter models (see for example Chapter 10 of Brandt et al. (2016)); our approach is to estimate the next vote from a random voter in the population. Since we cannot observe all voters, our model should generalize not only a "known" voter's future behavior, but *all* voters' future behavior.

In the second paradigm, *Representative Decisions*, we have a small number of "representative" voters; our goal is to estimate best responses to new decision problems for this group of representatives. This scenario is similar to multi-stakeholder decisions including organ allocation or public policy design: these decisions are made by a small number of representatives (e.g., experts in medicine or policy), who often have very limited time to express their preferences. As in *Population Modeling* we aim to estimate the next vote from a random expert—however in this paradigm, all voters are "known", i.e., in the training data.

Both voting paradigms can be represented as a machine learning problem: observed votes are "data", with which we select a best-fit model from a hypothesis class; these models make predictions about future votes.[10] Thus, we split all observed votes into a training set (for model fitting) and a test set (for evaluation). How we split votes into a training and test set is important: in *Representative Decisions* we aim to predict future decisions from a *known* pool of voters—so both the training and test set should contain votes from each voter. In *Population Modeling* we aim to predict future decisions from the entire voter population—so the training set should contain only some votes from some voters (i.e., "training" voters), while the test set should contain the remaining votes from training voters, and all responses from the non-training voters.

We propose several group indecision models, each of which is based on the models from Section ; please see Appendix **??** for more details.

`VMixture` Model. We first learn a best-fit indecision (sub)model for each training voter; the overall model gen-

[10]Several researchers have used techniques from machine learning for social choice (Doucette, Larson, and Cohen 2015; Conitzer et al. 2017; Kahng et al. 2019; Zhang and Conitzer 2019).

Model Name	*Represenatitives* (20)		*Population* (100)	
	Indecisive	*Strict*	*Indecisive*	*Strict*
2-Min-δ	-0.90/-0.88	-0.46/-0.47	-0.87/-0.88	-0.54/-0.52
2-Mixture	-0.87/**-0.86**	-0.45/-0.47	-0.87/-0.88	-0.53/-0.52
VMixture	-0.92/-0.90	-0.49/-0.51	-0.93/-0.94	-0.57/-0.56
Min-δ	-0.92/-0.90	-0.46/-0.48	-0.87/-0.87	-0.54/-0.53
Max-δ	-0.95/-0.90	-0.45/**-0.46**	-0.96/-0.95	-0.54/-0.52
Min-U	-0.96/-0.95	-0.52/-0.54	-0.98/-0.99	-0.58/-0.57
Max-U	-0.87/**-0.86**	-0.54/-0.54	-0.94/-0.94	-0.58/-0.57
Dom	-1.08/-1.07	-0.57/-0.58	-1.05/-1.06	-0.61/-0.60
MLP	-0.40/-1.55	-0.15/-0.85	-0.71/**-0.77**	-0.42/**-0.51**
Logit	-0.91/-0.88	-0.53/-0.54	-0.93/-0.94	-0.57/-0.56
Rand	-1.03/-1.00	N/A	-1.07/-1.07	N/A

Table 2: Average train-set and test-set LL per question (reported as "train/test") for *Representative Decisions* with 20 training voters, (left) and *Population Modeling* with 100 training voters (right), for both the *Indecisive* and *Strict* participant groups. The greatest test-set LL is highlighted for each column. For *Representatives*, the test set includes only votes from the representative voters; for *Population*, the test set includes all voters.

erates responses by first selecting a training voter uniformly at random, and then responding according to their submodel.

k-Mixture Model. This model consists of k submodels, each of which is an indecision model with its own utility vector $\boldsymbol{u}$ and threshold λ. The *type* of each submodel (Min/Max-δ, Min/Max-U, Dom) is itself a categorical variable. Weight parameters $\boldsymbol{w} \in \mathbb{R}^k$ indicate the importance of each submodel. This model votes by selecting a submodel from the softmax distribution[11] on $\boldsymbol{w}$, and responds according to the chosen submodel.

k-Min-δ Mixture. This model is equivalent to k-Mixture, however all submodels are of type Min-δ. We include this model since Min-δ is the most-common best-fit indecision model for individual participants (see §).

We simulate both the *Population Modeling* and *Representative Decisions* settings using various train/test splits of our survey data. For *Population Modeling* we randomly select 100 training voters; half of each training voter's responses are added to the test set, and half to the training set. All responses from non-training voters are added to the test set.[12]

For *Representative Decisions* we randomly select 20 training voters ("representatives"), and randomly select half of each voter's responses for testing; all other responses are used for training; all non-training voters are ignored.

For both of these settings we fit all mixture models (2-Mixture, 2-Min-δ, and VMixture), each individual indecision model from Section , and each each baseline model. Table 2 shows the training-set and test-set LL for each method, for both voting paradigms. Most indecision models achieve similar test-set LL, with the exception of Dom. In the *Representatives* setting, both mixture models and (non-mixture) indecision models perform well (notably, better than MLP. This is somewhat expected, as the *Representatives* setting uses very little training data, and complex ML approaches such as MLP are prone to overfitting—this is certainly the case in our experiments. In the *Population* setting the mixture models outperform individual indecision models; this is expected, as these mixture models have a strictly larger hypothesis class than any individual model. Unsurprisingly, MLP achieves the greatest test-set LL in the *Population* setting—yet provides no insight as to how these decisions are made.

[11] With the softmax distribution, the probability of selecting i is $e^{\boldsymbol{w}_i} / \sum_j e^{\boldsymbol{w}_j}$. We use this distribution for mathematical convenience, though it is straightforward to learn the distribution directly.

[12] Each voter in our data answers different questions, so all questions in the test set are "new."

Discussion

In many cases it is natural to feel indecisive, for example when voting in an election or buying a new car; people are especially indecisive when their choices have moral consequences. Importantly, there are many possible *causes* for indecision, and each conveys different meaning: I may be indecisive when voting for a presidential candidate because I feel unqualified to vote; I may be indecisive when buying a car because all options seem too similar. Using a small study, in Section we demonstrate that indecision cannot be interpreted as a "flipping a coin" to decide between alternatives. This violates a key assumption in the technical literature, and it complicates the task of selecting the *best* alternative for an individual or group. Indeed, defining the "best" alternative for indecisive agents depends on what indecision means.

These philosophical and psychological questions have become critical to computer science researchers, since we now use preference modeling and social choice to guide deployed AI systems. The indecision models we develop in Section and test in Section provide a framework for understanding why people are indecisive—and how indecision may influence expressed preferences when people are allowed to be indecisive (§), and when they are required to express strict preferences (§). The datasets collected in Study 1 (§) and Study 2 (§) provide some insight into the causes for indecision, and we believe other researchers will uncover more insights from this data in the future.

Several questions remain for future work. First, what are the causes for indecision, and what meaning do they convey? This question is well-studied in the philosophy and social science literature, and AI researchers would benefit from interdisciplinary collaboration. Methods for preference elicitation (Blum et al. 2004) and active learning (Freund et al. 1997) may be useful here.

Second, if indecision has meaning beyond the desire to "flip a coin", then what is the best outcome for an indecisive agent? ... for a group of indecisive agents? This might be seen as a problem of winner determination, from a perspective of social choice (Pini et al. 2011).

Acknowledgements

Dickerson and McElfresh were supported in part by NSF CAREER IIS-1846237, NSF CCF-1852352, NSF D-ISN #2039862, NIST MSE #20126334, NIH R01 NLM-013039-01, DARPA GARD #HR00112020007, DoD WHS #HQ003420F0035, DARPA Disruptioneering (SI3-CMD) #S4761, and a Google Faculty Research Award. Conitzer was supported in part by NSF IIS-1814056. This publication was made possible through the support of a grant (TWCF0321) from Templeton World Charity Foundation, Inc. to Conitzer, Schaich Borg, and Sinnott-Armstrong. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc.

Ethics Statement

Many AI systems are designed and constructed with the goal of promoting the interests, values, and preferences of users and stakeholders who are affected by the AI systems. Such systems are deployed to make or guide important decisions in a wide variety of contexts, including medicine, law, business, transportation, and the military. When these systems go wrong, they can cause irreparable harm and injustice. There is, thus, a strong moral imperative to determine which AI systems best serve the interests, values, and preferences of those who are or might be affected.

To satisfy this imperative, designers of AI systems need to know what affected parties really want and value. Most surveys and experiments that attempt to answer this question study decisions between two options without giving participants any chance to refuse to decide or to adopt a random method, such as flipping a coin. Our studies show that these common methods are inadequate, because providing this third option—which we call indecision—changes the preferences that participants express in their behavior. Our results also suggest that people often decide to use a random method for variety of reasons captured by the models we studied. Thus, we need to use these more complex methods—that is, to allow indecision—in order to discover and design AI systems to serve what people really value and see as morally permitted. That lesson is the first ethical implication of our research.

Our paper also teaches important ethical lessons regarding justice in data collection. It has been shown that biases can, and are, introduced at the level of data collection. Our results open the door to the suggestion that biases could be introduced when a participant's values are elicited under the assumption of a strict preference. Consider a simple case of choosing between two potential kidney recipients, A and B, who are identical in all aspects, except A has 3 drinks a week while B has 4. Throughout our studies, we have consistently observed that participants would overwhelmingly give the kidney to patient A who has 1 fewer drink each week, when forced to choose between them. However, when given the option to do so, most would rather flip a coin. An argument can be made here that the data collection mechanism under the strict-preference assumption is biased against patient B and others who drink more than average.

Finally, our studies also have significant relevance to randomness as a means of achieving fairness in algorithms. As our participants were asked to make moral decisions regarding who should get the kidney, one interpretation of their decisions to flip a coin is that the fair thing to do is often to flip a coin so that they (and humans in general) do not have to make an arbitrary decision. The modeling techniques proposed here differ from the approach to fairness that conceives random decisions as guaranteeing equity in the distribution of resources. Our findings about model fit suggest that humans sometimes employ random methods largely in order to avoid making a difficult decision (and perhaps also in order to avoid personal responsibility). If our techniques are applied to additional problems, they will further the discussion of algorithmic fairness by emphasizing the role of randomness and indecision. This advance can improve the ability of AI systems to serve their purposes within moral constraints.

Experiment Scenario: Organ Allocation. Our experiments focus on a hypothetical scenario involving the allocation of scarce donor organs. We use organ allocation since it is a real, ethically-fraught problem, which often involves AI or other algorithmic guidance. However our hypothetical organ allocation, and our survey experiments, are not intended to reflect the many ethical and logistical challenges of organ transplantation; these issues are settled by medical experts and policymakers. Our experiments do not focus on a realistic organ allocation scenario, and our results should not be interpreted as guidance for transplantation policy.

References

Agarwal, N.; Ashlagi, I.; Rees, M. A.; Somaini, P. J.; and Waldinger, D. C. 2019. Equilibrium Allocations under Alternative Waitlist Designs: Evidence from Deceased Donor Kidneys. Working Paper 25607, National Bureau of Economic Research. doi:10.3386/w25607. URL http://www.nber.org/papers/w25607.

Blackburn, S. 1996. Dilemmas: Dithering, Plumping, and Grief. In Mason, H. E., ed., *Moral Dilemmas and Moral Theory*, 127. Oxford University Press.

Blum, A.; Jackson, J.; Sandholm, T.; and Zinkevich, M. 2004. Preference elicitation and query learning. *Journal of Machine Learning Research* 5(Jun): 649–667.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of computational social choice*. Cambridge University Press.

Chang, R. 2002. The Possibility of Parity. *Ethics* 112(4): 659–688.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4831–4835.

Day, B.; Bateman, I. J.; Carson, R. T.; Dupont, D.; Louviere, J. J.; Morimoto, S.; Scarpa, R.; and Wang, P. 2012. Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of environmental economics and management* 63(1): 73–91.

DeShazo, J.; and Fermo, G. 2002. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and management* 44(1): 123–143.

Donagan, A. 1984. Consistency in Rationalist Moral Systems. *Journal of Philosophy* 81(6): 291–309. doi: jphil198481650.

Doucette, J. A.; Larson, K.; and Cohen, R. 2015. Conventional Machine Learning for Social Choice. In *AAAI*, 858–864.

Freedman, R.; Borg, J. S.; Sinnott-Armstrong, W.; Dickerson, J. P.; and Conitzer, V. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 103261.

Freund, Y.; Seung, H. S.; Shamir, E.; and Tishby, N. 1997. Selective sampling using the query by committee algorithm. *Machine learning* 28(2-3): 133–168.

Furnham, A.; Simmons, K.; and McClelland, A. 2000. Decisions concerning the allocation of scarce medical resources. *Journal of Social Behavior and Personality* 15(2): 185.

Furnham, A.; Thomson, K.; and McClelland, A. 2002. The allocation of scarce medical resources across medical conditions. *Psychology and Psychotherapy: Theory, Research and Practice* 75(2): 189–203.

Gangemi, A.; and Mancini, F. 2013. Moral choices: the influence of the do not play god principle. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society, Cooperative Minds: Social Interaction and Group Dynamics*, 2973–2977. Cognitive Science Society, Austin, TX.

Gerasimou, G. 2018. Indecisiveness, undesirability and overload revealed through rational choice deferral. *The Economic Journal* 128(614): 2450–2479.

Hare, R. M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

Kahng, A.; Lee, M. K.; Noothigattu, R.; Procaccia, A.; and Psomas, C.-A. 2019. Statistical foundations of virtual democracy. In *International Conference on Machine Learning*, 3173–3182.

Luce, M. F. 1998. Choosing to Avoid: Coping with Negatively Emotion-Laden Consumer Decisions. *Journal of Consumer Research* 24(4): 409–433.

Mattei, N.; Saffidine, A.; and Walsh, T. 2018. Fairness in deceased organ matching. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 236–242.

McElfresh, D. C.; and Dickerson, J. P. 2018. Balancing Lexicographic Fairness and a Utilitarian Objective with Application to Kidney Exchange. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

McIntyre, A. 1990. Moral Dilemmas. *Philosophy and Phenomenological Research* 50(n/a): 367–382. doi:10.2307/2108048.

Mochon, D. 2013. Single-option aversion. *Journal of Consumer Research* 40(3): 555–566.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A Voting-Based System for Ethical Decision Making. *AAAI 2018* .

Oedingen, C.; Bartling, T.; and Krauth, C. 2018. Public, medical professionals' and patients' preferences for the allocation of donor organs for transplantation: study protocol for discrete choice experiments. *BMJ open* 8(10): e026040.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pini, M. S.; Rossi, F.; Venable, K. B.; and Walsh, T. 2011. Incompleteness and incomparability in preference aggregation: Complexity results. *Artificial Intelligence* 175(7-8): 1272–1289.

Railton, P. 1992. Pluralism, Determinacy, and Dilemma. *Ethics* 102(4): 720–742. doi:10.1086/293445.

Scheunemann, L. P.; and White, D. B. 2011. The ethics and reality of rationing in medicine. *Chest* 140(6): 1625–1632.

Shapley, L. S.; and Shubik, M. 1974. *Game Theory in Economics: Chapter 4, Preferences and Utility*. Santa Monica, CA: RAND Corporation.

Sinnott-Armstrong, W. 1988. *Moral Dilemmas*. Blackwell.

Sobol', I. M. 1967. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 7(4): 784–802.

Soufiani, H. A.; Parkes, D. C.; and Xia, L. 2013. Preference elicitation for General Random Utility Models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 596–605.

Tversky, A.; and Shafir, E. 1992. Choice under conflict: The dynamics of deferred decision. *Psychological science* 3(6): 358–361.

Zakay, D. 1984. " To choose or not to choose": On choice strategy in face of a single alternative. *The American journal of psychology* 373–389.

Zhang, H.; and Conitzer, V. 2019. A PAC framework for aggregating agents' judgments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2237–2244.