TOWARDS COGNITIVELY-FAITHFUL DECISION-MAKING MODELS TO IMPROVE AI ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent AI trends seek to align AI models to learned *human-centric objectives*, such as personal preferences, utility, or societal values. Using standard preference elicitation methods, researchers and practitioners build models of human decisions and judgments, to which AI models are aligned. However, standard elicitation methods often fail to capture the cognitive processes behind human decision making, such as *heuristics* or simplifying *structured thought patterns*. To address this failure, we take an axiomatic approach to learning *cognitively faithful* decision processes from pairwise comparisons. Building on the literature analyzing cognitive processes that shape human decision-making, we derive a model class in which features are first *processed* with learned rules, then *aggregated* via a fixed rule, such as the Bradley-Terry rule, to produce a decision. This structured processing of information ensures that such models are realistic and feasible candidates to represent underlying human decision-making processes. We demonstrate the efficacy of this modeling approach by learning interpretable models of human decision making in a kidney allocation task, and show that our proposed models match or surpass the accuracy of prior models of human pairwise decision-making.

1 INTRODUCTION

Computational models of human cognition allow us to quantify and study the factors that influence our decisions, and when accurate, they help explain commonalities in human decision processes across different tasks (Gigerenzer et al., 2000; Crockett, 2016). This may also be why computational models to predict human behavior have recently found applications in personalized AI tools in various settings, e.g., healthcare, autonomous vehicles, and resource allocation, to increase users' confidence that AI will be aligned with their preferences (Ji et al., 2023; Kim et al., 2018; Noothigattu et al., 2018; 2019).

Despite promising use cases, current AI models of human decision-making typically do not attempt to faithfully replicate cognitive processes. Modern AI tools built using preference elicitation (Lee et al., 2019; Awad et al., 2018; Johnston et al., 2023; Freedman et al., 2020; Rafailov et al., 2023) and reinforcement learning (Ng et al., 2000; Christiano et al., 2017; Kaufmann et al., 2023) implicitly assume that a reward/objective model from a prespecified hypothesis class can accurately predict human decisions, but such tools are agnostic about how faithful these models are to cognitive processes. However, recent works highlight the lack of suitable hypothesis classes to capture human choices or their values (Stray et al., 2021; Boerstler et al., 2024; Leike et al., 2018). Unsuitable modeling classes introduce the possibility of inappropriate optimization formulations and *reward hacking* (Pan et al., 2022; Hubinger et al., 2019), and can lead to arbitrarily erroneous inferential models learned from human responses (Zhuang & Hadfield-Menell, 2020; Pan et al., 2022). Additionally, such misaligned models are limited in their explanatory power and *trustworthiness*. As suggested by Jacovi et al. (2021), misalignment between AI and human reasoning processes threatens *intrinsic trust* placed in the AI's decisions. These problems are further exacerbated when AI is used in high-stakes domains, including moral domains such as healthcare and sentencing (Sinnott-Armstrong et al., 2021; Kim et al., 2018), where stakeholders often expect an AI to justify its decision in a similar manner and to the same extent as humans (Lima et al., 2021; Keswani et al., 2025). Modeling done in prior moral decision-making contexts either favors simple model classes — e.g., linear models and decision trees (Lee et al., 2019; Noothigattu et al., 2018; Freedman et al., 2020), whose fit for human decision-making is highly context-specific (Ganzach, 2001; Zorman et al., 1997), or employs uninterpretable classes — e.g., neural networks or random forests (Wiedeman et al., 2020; Ueshima & Takikawa, 2024) — whose decision processes cannot be validated due to their uninterpretable designs. A qualitative study by Keswani et al. (2025) note this *process misalignment* as a crucial issue around AI-trust in morally-salient tasks like kidney transplant allocation, with one study participant expressing concern: "*they don't think like a human... what they might think should be ranked as priority, I may not.*"

Such human concerns highlight that, for certain real-world applications, building trustworthy AI tools requires accurately simulating human decision processes. However, human decision processes are often not captured by standard hypothesis classes. For example, consider humans' use of decision *heuristics*, defined by Gigerenzer & Gaissmaier (2011) as a strategy "*that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods*." Keswani et al. (2025) document several heuristics people use when deciding which patient should get an available kidney. Their study participants often used a form of *hiatus heuristic*, employing *thresholds* to isolate relevant patient feature information; e.g., some treated "number of dependents" as a binary (none or any), disregarding the actual *quantity*. Similarly, rather than contrast patients holistically, some participants used simple aggregation methods, such as the *tallying heuristic*, which simply counts the number of features favoring each patient to make their choices. While the role of such heuristics in human decision processes is well established (Gigerenzer & Gaissmaier, 2011; Shah & Oppenheimer, 2008; Gigerenzer et al., 2000), models learned from human decisions using standard hypothesis classes often don't capture or explain the heuristics people use to make decisions. For example, if a decision-maker uses the threshold-based heuristic mentioned above, linear models would fail to capture it, while neural networks or multivaritae monotonic regressors may learn the heuristic, but would fail to explain its role in the decision process. To explain or simulate human decision processes accurately, we need methods to learn and represent these heuristics (and other decision-making nuances). To that end, this paper explores hypothesis classes that more accurately capture the cognitive processes people use to make their decisions in pairwise comparison settings.

This work presents a learning framework for computational models of pairwise human decision-making, which is motivated by prior work on human cognition. We present natural decision-making axioms (sec. 2.2), and from them, we derive a class of feasible *cognitively faithful* decision-making models. Given any pairwise comparison between two options $x_1, x_2 \in \mathbb{R}^d$, we model the decision-maker's response to this comparison as a sequential two-stage process (sec. 2.1). The first stage captures the *editing rules* that the decision-maker employs over individual features $x_1^{(j)}, x_2^{(j)}$ to process/simplify/transform the information presented in feature $j \in [d]$. Each feature j may have its own decision rule, and these rules work towards transforming $x_i^{(j)}$ in a manner that reflects the feature's contribution to the decision-maker's final choice (e.g., whether the decision-maker thresholds, ignores, linearly transforms, or otherwise processes the feature). We allow this transformation to be *contextual* in nature, whereby the value of one or more features can influence the transformation used for another feature. Such *conditional transformations* capture feature interactions in our hypothesis class, increasing the models' expressiveness while ensuring that we still learn feature-level decision rules. The second stage captures the decision rule that aggregates the processed features to choose the *preferred option*. This aggregation rule can be as simple as the *tallying heuristic* (Gigerenzer et al., 2000) or as complicated as *Bradley-Terry aggregation* for probabilistic preferences (Bradley & Terry, 1952). Beyond the motivations from prior works on heuristic decision-making, we present a simple axiomatic basis of human decision-making processes that yields our two-stage model class (sec. 2.2). Moreover, more stringent assumptions yield special cases of interest, such as logistic regression, probit regression, and monotonic decision rules. Finally, we assess the empirical efficacy of our proposed framework over synthetic and real-world datasets in the kidney allocation domain (sec. 3), a domain where users have said cognitive process alignment is particularly important. We find that our method can learn models that explain the decision rules that people use for kidney allocation decisions, while ensuring that the learned model has similar or better predictive accuracy than other modeling approaches.

Related Work Methods for learning preferences from choices among options have been proposed in many contexts, such as recommender systems (Kalloori et al., 2018; Guo & Sanner, 2010; Chen & Pu, 2004), language model personalization (Rafailov et al., 2023; Jiang et al., 2024; Ziegler et al., 2019; Ouyang et al., 2022), and explanatory frameworks in social psychology (Lichtenstein & Slovic, 2006; Ben-Akiva et al., 2019; Charness et al., 2013). Recently, preference learning and elicitation of stakeholder preferences has flourished in the context of *human-centric AI* and AI alignment (Jiang et al., 2024; Ji et al., 2023; Capel & Brereton, 2023; Feffer et al., 2023; Kim et al., 2018; Johnston et al., 2023; Sinnott-Armstrong et al., 2021; Liscio et al., 2024; Lee et al., 2019). The usability of learned models in real settings relies on both their predictive accuracy and their alignment to humans' decision-making (Mukherjee et al., 2024; Keswani et al., 2025; Capel & Brereton, 2023; Lima et al., 2021). However, modeling strategies in most prior AI alignment work are agnostic about whether they actually capture human decision-making processes. Our work fills this gap by proposing modeling classes explicitly motivated by the decision rules people report using for pairwise comparisons. Like this work, Noothigattu et al. (2020) study the class of binary decision rules arising from pairwise comparisons. However, their axioms concern MLE estimation over data, and thus are *distribution sensitive*, whereas our axioms dictate laws as to how probabilistic preferences must behave *over their domain*. They also do not characterize model classes arising from their axioms, but rather study whether known classes satisfy them, making their work more taxonomically descriptive than prescriptive. Ge et al. (2024) continue this line of inquiry, again with axioms related to estimation from datasets.

There have also been other efforts to computationally model human decision-making. Bourgin et al. (2019) propose constructing models bounded by theoretical properties of human decision-making that are fine-tuned with real-world data, but they aim for accurate predictions, rather than cognitive fidelity or interpretable models. Peterson et al. (2021) use neural networks to implement and test cognitive models of participant choices for pairs of gambles. This analysis provides information on the *fit* of various cognitive models, but does not provide a mechanism to *learn* the decision-making process from an individual's data. Plonsky et al. (2017) and Payne et al. (1988) use psychological theories to select feature transformations or simulate prespecified heuristics to obtain gains in predictive power by curating the learning tasks to be more cognitive aligned with human decision processes. However, the applicability of these methods can be limited in settings where feature transformations are unknown *a priori* and vary across individuals. Our work also uses relevant works in psychology to identify appropriate decision rule assumptions, but we learn the feature heuristics/transformations from human decisions. Other works use neural networks alongside theory-driven cognitive models to simulate human decision-making (Fintz et al., 2022; Lin et al., 2022), but such models of learned decision processes are largely uninterpretable. Several studies also illustrate that simple heuristic-based models predict human responses well in classification (Holte, 1993; Şimşek & Buckmann, 2015; Brighton, 2006; Czerlinski et al., 1999; Dawes, 1979), supporting our goal of learning heuristics from decisions. Yet, the heterogeneity of heuristics used by different people makes it difficult to find a single model to accurately simulate the responses for an entire population. Our work addresses this issue by learning individual-level decision rules from data.

2 A Model of Cognitively-Faithful Decision-Making

In our setting, a decision-maker is presented with a pairwise comparison (x_1, x_2) between two options from the domain $\mathcal{X} \subseteq \mathbb{R}^d$, i.e. each with d descriptive features. For each $i \in [d]$, let $\mathcal{X}_i \subseteq \mathbb{R}$ denote the input space for feature i, such that $\mathcal{X} \doteq \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. The decision-maker's stochastic response function $H : \mathcal{X} \times \mathcal{X} \to [0, 1]$ denotes the probability of choosing the first option x_1 for any given pairwise comparison (x_1, x_2). To learn H from observed data, suppose we are given a dataset S containing the decision-maker's responses to N pairwise comparisons of the kind (x_1, x_2, r), where $r \in \{0, 1\}$, is the binary decision, with $r = 1$ if the decision-maker deterministically chose x_1, and 0 otherwise.

Given dataset S, we aim to learn a model $\hat{H}$ that accurately simulates the decision-maker's response function. Taking a general learning approach to this problem, given a hypothesis class of models $\mathcal{H}$, we can search for a model from $\mathcal{H}$ that minimizes the predictive loss over S. A common approach to learning $\hat{H}$ is to start with *standard* hypothesis classes, such as classes of linear functions, decision trees, or neural networks. However, as discussed earlier, the assumptions of these classes are likely not faithful to the cognitive processes people truly use to reach their own decisions (Keswani et al., 2025). As such, when using these classes, the resulting model can be difficult to validate and may lead to erroneous predictions in cases where the model class cannot capture the computation required to reach the "correct" decision. We aim to identify a hypothesis class that yields the most accurate model while faithfully representing humans' actual decision-making processes. Such models have the added benefit of being interpretable for human users. To come up with cognitively faithful computational models, we first discuss the computational properties commonly observed in human decision-making processes, and then use these properties to construct an appropriate hypothesis class.

2.1 Rule-Based Decision-Making Models

Prior works on cognitive models for human decision-making point to a reliance on *decision rules* and *heuristics* in comparative settings (Gigerenzer et al., 2000; Gigerenzer & Gaissmaier, 2011; Brandstätter et al., 2006; Kahneman & Tversky, 2013; Kahneman & Frederick, 2005). Based on this literature, we construct hypothesis classes that consist of hierarchical decision rules. Our proposed strategy models decision-making for pairwise comparison of options (x_1, x_2) as a two-step process. In the first step, the decision-maker processes each feature in x_1 and x_2 to *edit* or transform the information presented using this feature. In the second step, the decision-maker combines the edited information from all features to select the *dominant* option. Our hypothesis class will consist of functions that reflect this hierarchical process.

Editing Rules The decision rules or heuristic functions used in the first step are referred to as *editing rules*. Editing rules operate at the level of individual features of each element in the given pair and operationalize how a decision-maker processes the given feature value. Let $h^i_{\text{inn}} : \mathcal{X}_i \to \mathcal{X}'_i$ denote the editing rule for feature i, with $\mathcal{X}'_i$ the output domain post-editing. Examples of editing include feature simplification (e.g., zeroing out features considered irrelevant), transformation (e.g., log transformation for features with "diminishing returns" or scaling), or leaving the feature unchanged.

While these rules are often *structurally simple*, reflecting their use to reduce cognitive load (Gigerenzer & Gaissmaier, 2011), there can be significant heterogeneity in the editing used in different contexts. For example, prior works note *feature interactions*, where the editing rule used for the ith feature $x^{(i)}$ can depend on the values of other features of x (Keswani et al., 2025). To account for this, we will consider the choice of editing rule conditional on the *decision context* features, denoted by the values of features in a set $\omega \subseteq [d]$. On one extreme, $\omega = \emptyset$, implying that each feature is edited independently (i.e., no feature interactions), as ω grows, model complexity increases as conditional interactions involve more features, and the other extreme, $\omega = [d]$, implies that the choice of editing function for each feature in x can depend on the values of all other features. To account for this context ω, for any option $x \in \mathcal{X}$, we will denote the contextual editing rule operating over feature i by $h^{i,x^{\omega_i}}_{\text{inn}} : \mathcal{X}_i \to \mathcal{X}'_i$, where $\omega_i = \omega \setminus \{i\}$.

Remark 2.1 (Examples of Editing Rules)**.** *The use of editing rules is well-supported by literature in behavioral economics and social psychology. Montgomery (1983) describes the phase of "separating relevant information from less relevant information which can be discarded...," and Kahneman & Tversky (2013) argue that this allows for decision-making based only on the most essential information. In some settings, editing rules represent* feature importance assignment *(Payne, 1976; Shah & Oppenheimer, 2008; Gigerenzer et al., 2000). In other cases, editing reflects* feature transformations *to isolate information relevant to the task (Ajzen, 1996; Shah & Oppenheimer, 2008), e.g., thresholding to discretize features, or* log*-transformation to model diminishing returns (Tversky, 1972; Kubanek, 2017).*

Dominance Testing In the second step of the decision-making process, the edited features are compared across the two options (x_1, x_2) and then combined to reach the final decision on which option is the *dominant one*. We will refer to this second step as the *dominance testing rule*. Let $h_{\text{out}} : \mathcal{X}' \times \mathcal{X}' \to [0,1]$ denote the dominance test function, where $\mathcal{X}' = \mathcal{X}'_1 \times \cdots \times \mathcal{X}'_d$.

Remark 2.2 (Examples of Dominance Testing rules)**.** *Several dominance testing rules have been studied in prior literature, including the tallying up heuristic mentioned earlier (Czerlinski et al., 1999; Gigerenzer & Gaissmaier, 2011) and the prominent feature heuristic (choose the element favored by the most prominent feature for which there is non-zero difference in edited feature value) (Persson et al., 2022; Tversky et al., 1988). Alternately, one could create additive or probabilistic functions to capture various dominance testing rules observed in practice, e.g., Bradley-Terry aggregation (Bradley & Terry, 1952) (discussed in sec. 2.2).*

With this setup, we model a decision maker's response to any pairwise comparison as first applying editing functions $h^{i,x^{\omega_i}}_{\text{inn}}$ over each feature i for both options, given context ω and $\omega_i = \omega \setminus \{i\}$, and then combining the edited information to reach the final decision using the dominance testing rule h_{out}. Hence, our proposed hypothesis class contains such two-stage models, namely

$$\mathcal{H} \doteq \left\{ x_1, x_2 \mapsto h_{\text{out}} \left(\forall i \in [d], x_1^{(i)} \mapsto h^{i,x_1^{\omega_i}}_{\text{inn}}(x_1^{(i)}), x_2^{(i)} \mapsto h^{i,x_2^{\omega_i}}_{\text{inn}}(x_2^{(i)}) \right) \;\middle|\; h^{\cdot,\cdot}_{\text{inn}} \in \mathcal{H}_{\text{inn}}, h_{\text{out}} \in \mathcal{H}_{\text{out}} \right\} .$$

The properties that characterize classes $\mathcal{H}_{\text{inner}}$ and $\mathcal{H}_{\text{outer}}$ are discussed in the next section.

2.2 Axiomatic Characterization of Decision Rules

The above-described two-stage process is motivated by extensive decision-making literature. An additional compelling motivation for this hypothesis class comes from an axiomatic characterization of human decision processes for pairwise comparisons. The axioms below describe simple properties expected to be satisfied in an ideal comparative choice model (independent of its functional form).

Definition 2.3 (Axioms of Binary Choice)**.** *The following axioms describe a binary preference function $H(x_1, x_2) : \mathcal{X}^2 \to \mathcal{Y}$. Unless otherwise stated, each must hold* for all $x_1, x_2, x_3 \in \mathcal{X}$.
1. ***Complementarity****: The order in which two options are presented should not impact the option that is eventually selected. Formally, we require that $H(x_1, x_2) = 1 - H(x_2, x_1)$.*

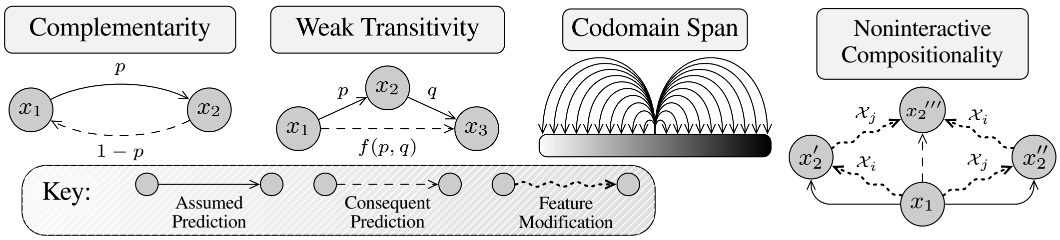

Figure 1: Visualization of the axioms of def. 2.3. Solid arrows represent assumed choice probabilities, dashed arrows represent choice probabilities implied by an axiom, and in *noninteractive compositionality*, dotted arrows represent feature modifications. Arrows may be labeled with choice probabilities.

2. ***Weak Transitivity*** *(WT): The comparison $H(x_1, x_3)$ can be expressed as a function of $H(x_1, x_2)$ and $H(x_2, x_3)$, i.e., for some $f : [0,1]^2 \to [0,1]$, it holds that $H(x_1, x_3) = f(H(x_1, x_2), H(x_2, x_3))$.*
3. ***Codomain Span****: For any $p \in (0,1)$ and $x_1 \in \mathcal{X}$, there exists $x_2 \in \mathcal{X}$ such that $H(x_1, x_2) = p$.*
4. ***Noninteractive Compositionality*** *(NC): Suppose some $x_1 \in \mathcal{X}$, and $x_2', x_2'', x_2''' \in \mathcal{X}$ such that x_2' and x_2'' are obtained by changing different features of x_1, and x_2''' is produced by applying both changes to x_1. We then require that $H(x_1, x_2''')$ can be computed from $H(x_1, x_2')$ and $H(x_1, x_2'')$.*
5. ***Conditionally Interactive Compositionality*** *(CIC): Given a set of* condition features *$\omega \subseteq [d]$, suppose some $x_1 \in \mathcal{X}$, and $x_2', x_2'', x_2''' \in \mathcal{X}$ such that x_2' and x_2'' are obtained by changing different features (neither in ω) of x_1, and x_2''' is produced by applying both changes to x_1. We then require that $H(x_1, x_2''')$ can be computed from $H(x_1, x_2')$ and $H(x_1, x_2'')$.*

These axioms are visualized in fig. 1. Complementarity ensures that the order in which options are presented should not matter, and predicted probabilities sum to 1. Weak transitivity requires that comparisons between (x_1, x_2) and (x_2, x_3) carry all the information about the items x_1, x_2, x_3 to also compare (x_1, x_3), i.e., the first two pairwise comparisons suffice to "complete the triangle" of all three pairwise comparisons. Codomain span is a technical condition, as to specify the decision rule for all probabilities, we must ensure that all probabilities arise in comparisons over the domain $\mathcal{X}$. Thm. 2.4 item 2 shows that these properties work together to induce a strong form of transitivity in probabilistic predictions.

Axioms 4 & 5 characterize the extent to which different features interact in the decision process. *Noninteractive compositionality* encodes a form of *feature independence*, requiring that the impact of changing two different features is compositional, which is less restrictive than assuming linearity, but restricts interactions between features in a similar manner. Essentially, this axiom characterizes a *generalized additive model* (GAM) (Hastie, 2017), as it dictates how information is synthesizes *across features*. *Conditional interactivity* then generalizes this idea, allowing for non-additive *conditional feature interactions*. We propose this axiom to account for a larger class of models that consider the decision context, where the impact of changing two features is additively compositional, *except if* one of the features is a *condition feature* (part of the *context* ω described in sec. 2.1). Axioms 1–3 are prescriptive, while 4–5 are not assumed to be universal, but rather characterize a class of simple decision-making rules.

All of our axioms are simple statements about concrete judgments over pairwise choices (rather than more abstract properties, like estimation or a specific decision mechanism). Furthermore, they are quite weak: WT assumes that transitive relationships *exist*, and NC and CIC assume that the impact of multiple independent feature modifications *can be computed*, without specifying any *particular* relationship. Note that we do not claim that our axioms characterize rationality, but rather an ideal binary choice process *in the absence of cognitive biases*; limitations of this assumption are discussed in sec. 4. We next show that the decision-making processes that satisfy these axioms follow the two-stage process outlined in sec. 2.1.

Theorem 2.4 (Axiomatic Factoring Characterization)**.** *We now explore the consequences of our axioms on characterizing the response function $H(\cdot,\cdot)$ of a decision-maker.*
1. ***Atomic Model*** *Suppose axiom 1, and also that $\mathcal{X}$ is a countable domain. Then there exists an* atomic rule *$h_{inn} : \mathcal{X} \to \mathbb{R}$ and $h_{out} : \mathbb{R} \to [0,1]$ s.t. $H(\cdot,\cdot)$ can be factored as*

$$H(x_1, x_2) = h_{out}\left(h_{inn}(x_1) - h_{inn}(x_2)\right) \quad .$$

2. *$\sigma(\cdot)$-**Transitivity** Suppose axioms 2–3 hold for some* continuous *transitivity law $f(\cdot,\cdot)$. Then there exists a symmetric continuous random variable with full support and CDF $\sigma(\cdot)$ s.t.*

$$H(x_1, x_3) = \sigma\left(\sigma^{-1}(H(x_1, x_2)) + \sigma^{-1}(H(x_2, x_3))\right) \quad .$$

Moreover, there exists some atomic rule $h_{inn}: \mathcal{X} \to \mathbb{R}$ *such that* $H(\cdot,\cdot)$ *can be factored as*

$$H(x_1, x_2) = \sigma\left(h_{inn}(x_1) - h_{inn}(x_2)\right) \quad .$$

3. ***Unconditional Factor Model**: Suppose as in item 2, and also axiom 4 (NC). Then there exist* inner functions $h^i_{inn}: \mathcal{X}_i \to \mathbb{R}$ *for* $i \in [d]$ *such that the atomic rule* $h_{inn}(\cdot)$ *and decision rule* $H(\cdot,\cdot)$ *factor as*

$$h_{inn}(x) = \sum_{i=1}^{d} h^i_{inn}(x^{(i)}) \quad , \quad H(x_1, x_2) = h_{out}\left(\sum_{i=1}^{d} h^i_{inn}(x_1^{(i)}) - h^i_{inn}(x_2^{(i)})\right) \quad .$$

4. ***Conditional Factor Model**: Suppose as in item 2, and also axiom 5 (CIC). Then there exist* conditional inner functions $h^{i,\cdot}_{inn}: \mathcal{X}_i \to \mathbb{R}$ *for* $i \in [d]$ *and* condition features *(context)* $\omega \subseteq [d]$, $\omega_i = \omega \setminus \{i\}$, *s.t.*

$$H(x_1, x_2) = h_{out}\left(\sum_{i=1}^{d} h^{i,x_1^{\omega_i}}_{inn}(x_1^{(i)}) - h^{i,x_2^{\omega_i}}_{inn}(x_2^{(i)})\right) \quad .$$

Item 1 shows that axiom 1 suffices to reduce binary choices to a function of a difference of atomic predictions for continuous transitivity laws $f(\cdot,\cdot)$, i.e., to $H(x_1, x_2) = h_{\text{out}}(h_{\text{inn}}(x_1) - h_{\text{inn}}(x_2))$, thus recovering the two-stage modeling process described in sec. 2.1. However, there is little structure beyond this: Such a factoring *exists*, but it is by no means *unique*, and it may well be *intransitive*. Item 2 then imposes weak transitivity through axiom 2, additional *continuity structure* through axiom 3, and paired with continuity of the transitivity law $f(\cdot,\cdot)$, we may conclude that $h_{\text{out}}(\cdot)$ is a sigmoid function (CDF). Axioms 4–5 are then needed to control feature interactions, and dictate a two-stage model structure wherein each feature is processed individually or conditionally.

Intuitively, with σ-transitivity, we require that transitive probabilities can be computed via addition within the sigmoid. Therefore, transitivity along chains of probabilities above $\frac{1}{2}$ grow ever closer to 1 as $\sigma(\cdot)$ is applied to a growing sum. In the parlance of generalized linear models, σ^{-1} plays the role of a link function. Bradley-Terry models, such as logistic regressors, employ the logistic sigmoid, but other sigmoids, such as the Gaussian CDF (with probit link function) also see use, e.g., in GLMs (MacCullagh & Nelder, 1989).

We now describe *strong assumptions* that may only be appropriate in specific real-world domains but nonetheless produce specific properties for the hypothesis class of editing function $\mathcal{H}_{\text{inner}}$.

Definition 2.5 (Assumptions on Binary Preference Models)**.** *We state assumptions on H that are suitable for discrete or continuous domains. These discrete properties must apply to all* $x_1, x_2 \in \mathcal{X}$.
1. ***σ-Linearity**: There exists some* $\boldsymbol{w} \in \mathbb{R}^d$ *such that* $\sigma^{-1} \circ H(x_1, x_2) = \boldsymbol{w} \cdot (x_1 - x_2)$.
2. ***Monotonicity**: If* $x_1 \preceq x_2$, *then* $H(x_1, x_2) \leq \frac{1}{2}$.

We now combine these stronger assumptions with the results of the axiomatic analysis of thm. 2.4. Recall that our basic axioms imply the factoring $\mathcal{H} = \{x_1, x_2 \mapsto \sigma\left(h_{\text{inn}}(x_1) - h_{\text{inn}}(x_2)\right) \,|\, h_{\text{inn}} \in \mathcal{H}_{\text{inn}}\}$.

Theorem 2.6 (Axiomatic Models)**.** *Suppose axioms 1–4. The following then restrict* $\mathcal{H}_{inn}$ *or each* $\mathcal{H}^{(i)}_{inn}$.
1. *Suppose* σ-linearity. *Then* $\mathcal{H}^{(i)}_{inn} = \{x \mapsto wx | w \in \mathbb{R}\}$, *thus* $\mathcal{H} = \{x_1, x_2 \mapsto \sigma(\boldsymbol{w} \cdot (x_1 - x_2)) \,|\, \boldsymbol{w} \in \mathbb{R}^d\}$.
2. *Suppose* monotonicity. *Then* $\mathcal{H}^{(i)}_{inn} = \{h : \mathbb{R} \to \mathbb{R} \,|\, x \leq y \implies h(x) \leq h(y)\}$.
3. ***Multivariate Monotonic Models**: If we assume monotonicity but relax noninteractive compositionality, then* $\mathcal{H} = \{x_1, x_2 \mapsto \sigma(h(x_1) - h(x_2)) \,|\, h : \mathcal{X} \to \mathbb{R}$ *s.t.* $\vec{x} \preceq \vec{y} \implies h(\vec{x}) \leq h(\vec{y})\}$.
4. ***Conditional GAM Tree** If we relax NC to CIC, then* $h_{inn}(\cdot)$ *may be represented as a hybrid model of a* decision tree */ GAM model, starting with a tree over* $\mathcal{X}_\omega$, *where each leaf contains a* GAM *over* $\mathcal{X}_{\setminus\omega}$. *Formally, we have* $\mathcal{H}_{inn} = \left\{\sum_{i=1}^{d-|\omega|}\left(T_i(x^\omega)\right)\left((x^{\setminus\omega})^{(i)}\right) \,\middle|\, T \in \mathbb{R}^{|\omega|} \to (\mathbb{R} \to \mathbb{R})^{d-|\omega|}\right\}$.

From this perspective, we derive a few valuable insights: *Logistic regression* with nonnegative weights is abstractly characterized by Bradley-Terry aggregation ($\sigma(u) = \text{logistic}(u)$) and linearity (and implicitly implies noninteractivity and monotonicity). If linearity is relaxed to monotonicity, we obtain the class of univariate monotonic models, and subsequently when noninteractivity is relaxed, we obtain univariate monotonic models with conditional interactions. Similarly, *probit regression* is characterized by taking $\sigma(u) = \Phi(u)$, i.e., the Gaussian CDF function, and linearity.

Extending Pairwise Comparisons to n-Way Choice

We focus on binary preference elicitation because it is a well-studied task in the decision-making literature. To extend our binary preference framework to learn probabilistic rankings over n items, we use the *choice axiom* of Luce et al. (1959): *Introducing additional items should not change the probability ratio of choosing* x_1 *to choosing* x_2, Without WT, there are $\binom{n}{2}$ degrees of freedom (DoF) to complementary binary preferences, and a probabilistic ranking that accords with these binary preferences in general does not exist (e.g., preferences over rocks-paper-scissors cannot accord with the choice axiom). However, with WT, there are only $n-1$ DoF, as a tree of predictions spanning n items can be completed via transitivity.

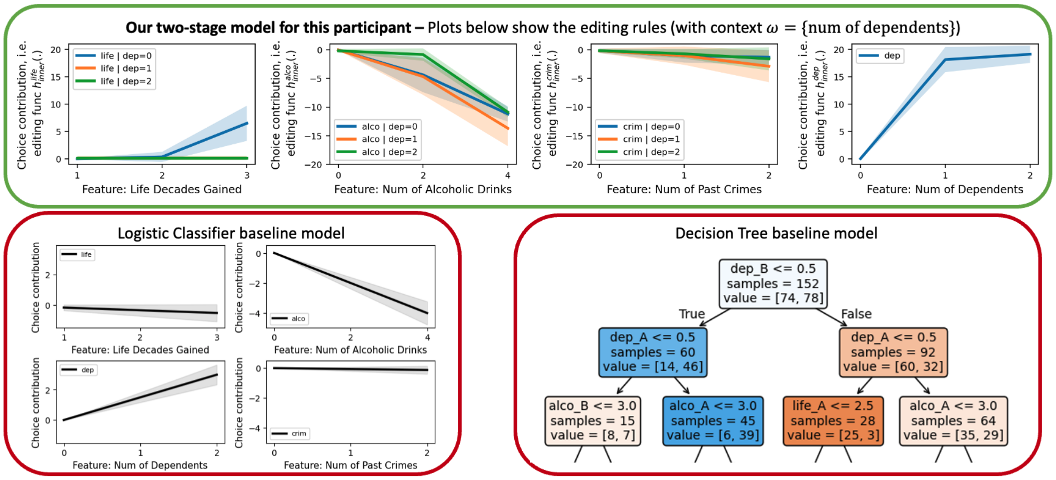

Figure 2: Interpretable models learned for P4 of Study One. For our two-stage model and logistic classifier, we plot the contributions of each value of each feature to the final choice i.e., we plot $h_{\text{inn}}(\cdot)$, which by thm. 2.6 item 1 is linear for the logistic model). We also show the decision tree, truncated to the first three layers. Note the differences in model interpretation between these strategies, with our model providing insight into the heuristics used by P4.

Theorem 2.7 (Proportionality of Choice). *Suppose a probabilistic decision rule over $n \geq 3$ items that obeys the* choice axiom. *Then $H(\cdot, \cdot)$ admits a σ-transitive factoring with logistic $\sigma(\cdot)$.*

Thm. 2.7 uniquely characterizes the Bradley-Terry model, but is only interesting insofar as the model is believable. Alternative models of n-way decision making are also well-studied, such as noisy observation, where utility + noise is observed for each option, and the largest observation wins, resulting in a probability distribution over outcomes. In particular, the Bradley-Terry model arises (non-uniquely) from homoskedastic Gumbel noise, but of course the Thurstone–Mosteller model (Thurstone, 1927; Mosteller, 1951) arises from homoskedastic Gaussian noise (hence the probit link function), and from this perspective, σ-transitivity can be generalized to n-way decision models whenever the distribution with CDF σ decomposes as the difference of two i.i.d. random variables, i.e., is in the *difference-form decomposable* (DFD) family, although such decompositions *are not always unique* (Carnal & Dozzi, 1989; Ewerhart & Serena, 2025).

Overall, with reasonable domain-specific assumptions, we recover other standard modeling classes from our two-stage modeling framework, highlighting the expressive power of our modeling class. Importantly, assumptions like linearity and monotonicity are domain-dependent and need to be qualitatively justified for the problem context. Classes, like logistic regressors, that implicitly encode these assumptions may not generate cognitively-faithful models when the human decision process violates these assumptions. Instead, we based our modeling class on generic axioms of human decision processes, which can be further constrained by domain-specific assumptions as necessary.

Additionally, our axiomatic approach significantly constrains the set of feasible models, thus simplifying the model class. For instance, suppose we are given a pairwise comparison $(x_1, x_2) \in \mathbb{R}^{2\times d}$. Assuming feature independence, a generic supervised learning approach would have at least $2d$ parameters. In comparison, due to the *complementarity* axiom, our approach constructs d editing functions, halving the parameter space, and implicitly enforcing symmetry over the two presented options, which is more cognitively plausible than models that process both options independently, essentially repeating the logic twice.

3 Empirical Analysis on Kidney Allocation Data

We test our modeling strategy to learn moral judgment processes for kidney allocation, where prior works study moral judgments regarding the allocation of kidneys to patients based on their medical attributes (e.g., transplant outcomes) and non-medical attributes (e.g., dependents and lifestyles) (Boerstler et al., 2024; Keswani et al., 2025; Chan et al., 2024). We assess with both real-world and synthetic data.

Real-world dataset We use the dataset of Boerstler et al. (2024), which spans two studies where participants were presented with several kidney allocation scenarios (15 participants in Study One and 40 in Study Two). In both studies, each kidney allocation scenario comprises a pairwise comparison between two patients (see fig. 4 for an example). Participants were instructed to choose which patient should be given an available kidney. In Study One, the patient features presented are the patient's *number of dependents*, projected *life years gained* from the transplant (LYG), *alcoholic drinks per day*, and *number of crimes committed*. In Study Two, the patient features are the patient's *number of elderly dependents*, LYG, *years waiting* for the transplant, *weekly work hours post-transplant*, and *obesity level*. We present further details in appx. B.

Synthetic dataset We also created a synthetic dataset representing multiple simulated decision-makers. This dataset contains pairwise comparisons of hypothetical kidney patients described using four features: *number of dependents*, LYG, *years waiting*, and *number of crimes* committed. The goal is to simulate decision-makers who explicitly use the heuristics observed in prior studies to assess how well our model and other baselines recover these heuristics. The heuristics are taken from Keswani et al. (2025), who provide user-reported qualitative accounts of decision rules people use for kidney allocation decisions. Using their observations, we create five simulated decision-makers, DM1–DM5, each using different decision rules over the presented patient features. For instance, DM1 decides between A and B with the following process: (a) assign 1 point to the patient with a higher LYG; (b) assign 1 point to each patient with dependents; (c) assign 1 point to each patient on the waiting list for >6 years; (d) add $\mathcal{N}(0,1)$ noise (i.e., a homoskedastic Thurstone-Mosteller process) to the difference of patient scores, and choose A if difference> 0 and B otherwise (DM1's process is mathematically described in fig. 3). Other simulated decision-makers (DM2–DM5) also use various heuristics, such as *thresholding*, *diminishing returns*, and *tallying* (described in appx. B). Our dataset contains 1000 pairwise comparisons from each simulated DM.

Methodology and Baselines We compare our model to several learning approaches, namely Bradley-Terry and Drift Diffusion models from cognitive science literature, and common supervised learning approaches, such as logistic and elastic-net classifiers, SVM, GAM with spline terms, decision trees, multi-layer perception (MLP), k-NN, and random forests. Drift Diffusion is the only model we consider that requires reaction times per scenario, thus it is not run for the simulated DMs. We use logistic models and decision tree models as the "interpretable" methods to compare our model's interpretability to. Since moral judgments vary substantially across individuals, all experiments operate over individual-participant-level data. For each decision-maker (real or simulated), we use a 70-30 train-test split, and report test accuracy over 20 repetitions. For our framework, we minimize the predictive loss to learn the editing functions $h_{\text{inn}}^{\cdot;\cdot}$ (which assign score$\in \mathbb{R}$ to each feature value), constraining all $h_{\text{inn}}^{\cdot;\cdot}$ to be monotonic (since all features in our dataset can be seen to impact the final choice monotonically) We implement two variants of this framework: (A) with cross-entropy loss and $\sigma(x) = (1 + e^{-x})^{-1}$, and (B) with hinge loss and σ as the identity function. The first variant is aligned with the probabilistic framework of thm. 2.4, while the second framework is better suited to assessing our learned model on the "hard classification" metric of 0-1 predictive accuracy. Context ω is limited to one feature for the real-world datasets, chosen using cross-validation, and $\emptyset$ for the synthetic dataset. Appx. B provides additional implementation details.

Results Across all datasets, we observe that our cognitively-motivated models achieve high accuracy and provide deeper insight into decision-making processes than other baselines.

Model	Average Accuracy (stddev) Study One	Study Two	Simulated
Cognitive models			
Drift-Diffusion	.89 (.05)	.88 (.05)	–
Bradley-Terry	**.90** (.06)	.78 (.06)	.77 (.06)
Supervised learning models			
Logistic Clf	**.90** (.06)	.89 (.05)	.85 (.07)
Elastic Net	.89 (.04)	.88 (.05)	.85 (.07)
SVM	.89 (.06)	.89 (.05)	.85 (.07)
GAM	.87 (.09)	.84 (.11)	.88 (.08)
Decision Tree	.83 (.06)	.79 (.06)	.82 (.11)
k-NN	.85 (.06)	.82 (.05)	.79 (.08)
MLP	.89 (.05)	.86 (.06)	.87 (.08)
Random Forest	.86 (.05)	.85 (.04)	.87 (.08)
Our Models			
cross-entropy loss	**.90** (.06)	**.90** (.05)	**.89** (.10)
hinge loss	**.90** (.06)	.89 (.06)	**.89** (.08)

Table 1: Performance statistics of all models on kidney allocation datasets.

Aggregated Performance. The mean accuracy of each model over all participants is shown in tbl. 1. Overall, on both real and synthetic data, our framework produces models *at least as accurate* as all baselines. Individual-level performance is presented in appx. B. Additionally, editing functions h_{inn} learned for the decision-makers provide insight into how they process the input features to reach their final decision. We demonstrate this interpretability through case studies of the editing functions learned for the simulated decision-maker DM1 and a real-world participant P4 from Study One.

Participant P4 in Study One. The two-stage model learned for P4 (w/ hinge loss) is illustrated in fig. 2 (top row). The editing function plots show the following nuances of their decision-making process for pairwise comparisons of kidney patients: (a) *number of dependents* and *number of alcoholic drinks* are the two most important features, as reflected by the large choice contributions of these features; (b) *number of past crimes* is considered mostly irrelevant by this participant; (c) *life decades gained* is relevant only when the patient has zero dependents, indicating a conditional interaction between these two features; (d) for *number of dependents*, a difference of 1 vs 0 dependents is much more significant to the final choice than a difference of 2 vs 1 dependents — approximately a

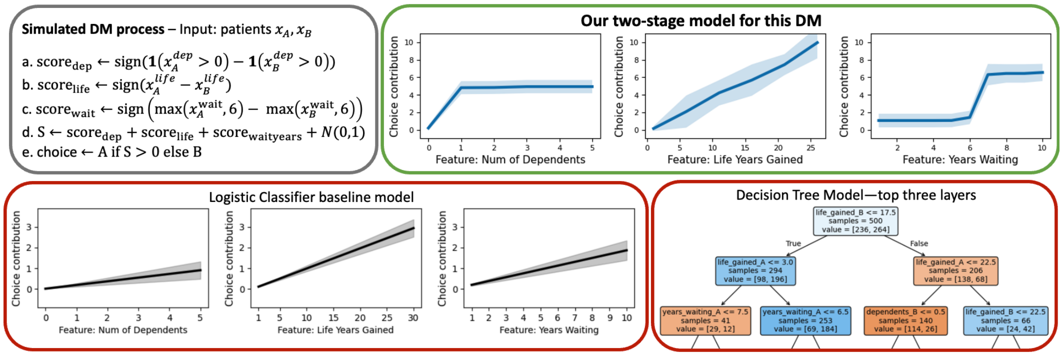

Figure 3: Models learned for the simulated decision-maker DM1. On the top left, we present the mathematical description of the DM1's process. Once again, we present each model's interpretation of the process used by the decision-maker. Observe that our two-stage model most accurately captures the simulated process.

threshold decision rule such that any non-zero number of dependents contributes equally; (e) similarly for *life decades gained* and *number of alcoholic drinks*, for certain conditions on *number of dependents*, the participant employs threshold decision rules. Our model is able to learn all of these decision-making nuances with an accuracy of 0.78 ($\pm$ 0.05), but a trained logistic classifier (bottom-left in fig. 2) has a lower accuracy of 0.76 ($\pm$ 0.04) and only uncovered points (a) and (b) above, while a trained decision-tree model (bottom-right in fig. 2) has an accuracy of 0.70 ($\pm$ 0.03) and only uncovered points (a), (c), and (d).

Simulated DM1. Fig. 3 (top right) shows how our model recovers the rules used by DM1. E.g., for the feature *number for dependents*, $h_{\text{inn}}^{\#\text{deps}}$ captures the increase in choice contribution when this feature has a non-zero value, and for the *years waiting* feature, $h_{\text{inn}}^{\text{wait yrs}}$ learns the increased contribution of values exceeding 6 years. Both successfully represent the threshold functions used in DM1. Note that neither of these observations can be easily inferred from logistic or decision tree models (bottom row of fig. 3). Our model is also more accurate (0.76 $\pm$ 0.02) than logistic (0.74 $\pm$ 0.02) and decision tree (0.68 $\pm$ 0.04) models. Appx. B shows similar results for other simulated decision-makers. Overall, our models provide a better understanding of the decision-making processes, without sacrificing predictive accuracy.

4 Discussion, Limitations, and Future Work

This work provides a modeling strategy to learn human decision processes from pairwise comparisons. The primary goal of this strategy is *fidelity*, such that best-fit models from these classes are more faithful to humans' actual decision-making than standard hypothesis spaces. Our theoretical analysis demonstrates that these classes emerge from natural decision-making axioms, and our empirical analysis shows that greater *fidelity* would lead to greater model *accuracy* in some cases. The advantages of cognitively-faithful models will depend on the application. In applications like online recommender systems, it may be sufficient to predict human behavior accurately, without simulating their actual decision processes. However, cognitively-faithful models would be desirable in other high-stakes AI domains (e.g., healthcare and sentencing), where the interpretability and alignment of AI's decisions with the users is pivotal in establishing trust in AI (Jacovi et al., 2021). Cognitively-faithful models would also be preferable in domains with no "ground truth" (like the kidney allocation domain), where model validation requires the user to understand and evaluate the model's decision process. Lastly, cognitively-faithful models are essential for any use of AI in moral domains, since stakeholders expect an AI to justify its moral decisions in a similar manner and to the same extent as humans (Lima et al., 2021).

As for limitations, the interpretability features of our framework need further validation through real-world user studies. While this work provides a proof-of-concept for a computational framework to learn from human decisions, future user studies can help assess the extent to which users understand and trust the learned two-stage models. Our axioms characterize the "ideal" decision-making process of an individual; i.e., the one they would prefer to follow in the absence of any internal/external constraints and/or the one that describes their judgments about what "should be done." However, human decisions are known to deviate from the ideal in many ways, e.g., well-known transitivity and complementarity violations (Tversky et al., 1990), among others. Future work can explore ways to extend our two-stage hypothesis classes to quantify deviations of one's "ideal" judgments from their implemented choices.

References

Icek Ajzen. The social psychology of decision making. *Social psychology: Handbook of basic principles*, pp. 297–325, 1996.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Moshe Ben-Akiva, Daniel McFadden, Kenneth Train, et al. Foundations of stated preference elicitation: Consumer behavior and choice-based conjoint analysis. *Foundations and Trends® in Econometrics*, 10 (1-2):1–144, 2019.

Kyle Boerstler, Vijay Keswani, Lok Chan, Jana Schaich Borg, Vincent Conitzer, Hoda Heidari, and Walter Sinnott-Armstrong. On the stability of moral preferences: A problem with computational elicitation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 156–167, 2024.

David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pp. 5133–5141. PMLR, 2019.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Eduard Brandstätter, Gerd Gigerenzer, and Ralph Hertwig. The priority heuristic: making choices without trade-offs. *Psychological review*, 113(2):409, 2006.

Henry Brighton. Robust inference with simple cognitive models. In *AAAI spring symposium: Between a rock and a hard place: Cognitive science principles meet AI-hard problems*, pp. 17–22, 2006.

Tara Capel and Margot Brereton. What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–23, 2023.

H Carnal and M Dozzi. On a decomposition problem for multivariate probability measures. *Journal of multivariate analysis*, 31(2):165–177, 1989.

Lok Chan, Walter Sinnott-Armstrong, Jana Schaich Borg, and Vincent Conitzer. Should responsibility affect who gets a kidney? *Responsibility and Healthcare*, pp. 35, 2024.

Gary Charness, Uri Gneezy, and Alex Imas. Experimental methods: Eliciting risk preferences. *Journal of economic behavior & organization*, 87:43–51, 2013.

Li Chen and Pearl Pu. Survey of preference elicitation methods. 2004.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Molly J Crockett. Computational modeling of moral decisions. In *The social psychology of morality*, pp. 71–90. Routledge, 2016.

Jean Czerlinski, Gerd Gigerenzer, and Daniel G Goldstein. How good are simple heuristics? In *Simple heuristics that make us smart*, pp. 97–118. Oxford University Press, 1999.

Robyn M Dawes. The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 1979.

Christian Ewerhart and Marco Serena. On the (im-)possibility of representing probability distributions as a difference of i.i.d. noise terms. *Mathematics of Operations Research*, 50(1):390–409, 2025.

Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. From preference elicitation to participatory ml: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 38–48, 2023.

Matan Fintz, Margarita Osadchy, and Uri Hertz. Using deep learning to predict human decisions and using cognitive models to explain deep learning models. *Scientific reports*, 12(1):4736, 2022.

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

Yoav Ganzach. Nonlinear models of clinical judgment: Communal nonlinearity and nonlinear accuracy. *Psychological Science*, 12(5):403–407, 2001.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI alignment from human feedback. *arXiv preprint arXiv:2405.14758*, 2024.

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62 (2011):451–482, 2011.

Gerd Gigerenzer, Peter M Todd, the ABC Research Group, et al. *Simple heuristics that make us smart*. Oxford University Press, 2000.

Shengbo Guo and Scott Sanner. Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 289–296. JMLR Workshop and Conference Proceedings, 2010.

Trevor J Hastie. Generalized additive models. *Statistical models in S*, pp. 249–307, 2017.

Robert C Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11:63–90, 1993.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635, 2021.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models. *arXiv preprint arXiv:2406.11191*, 2024.

Caroline M Johnston, Patrick Vossler, Simon Blessenohl, and Phebe Vayanos. Deploying a robust active preference elicitation algorithm on mturk: Experiment design, interface, and evaluation for covid-19 patient prioritization. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–10, 2023.

Daniel Kahneman and Shane Frederick. A model of heuristic judgment. *The Cambridge handbook of thinking and reasoning*, 267:293, 2005.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Saikishore Kalloori, Francesco Ricci, and Rosella Gennari. Eliciting pairwise preferences in recommender systems. In *Proceedings of the 12th acm conference on recommender systems*, pp. 329–337, 2018.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10, 2023.

Vijay Keswani, Vincent Conitzer, Walter Sinnott-Armstrong, Breanna K Nguyen, Hoda Heidari, and Jana Schaich Borg. Can ai model the complexities of human moral decision-making? a qualitative study of kidney allocation decisions. *arXiv preprint arXiv:2503.00940*, 2025.

Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B Tenenbaum, and Iyad Rahwan. A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 197–203, 2018.

Jan Kubanek. Optimal decision making and matching are tied through diminishing returns. *Proceedings of the National Academy of Sciences*, 114(32):8499–8504, 2017.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–35, 2019.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Sarah Lichtenstein and Paul Slovic. The construction of preference: An overview. *The construction of preference*, 1:1–40, 2006.

Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–17, 2021.

Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. Predicting human decision making in psychological tasks with recurrent neural networks. *PloS one*, 17(5):e0267907, 2022.

Enrico Liscio, Luciano Cavalcante Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Value preferences estimation and disambiguation in hybrid participatory systems. *CoRR*, abs/2402.16751, 2024. doi: 10.48550/ARXIV.2402.16751. URL https://doi.org/10.48550/arXiv.2402.16751.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

P MacCullagh and JA Nelder. Generalized linear models. 1989.

Henry Montgomery. Decision rules and the search for a dominance structure: Towards a process model of decision making. In *Advances in psychology*, volume 14, pp. 343–369. Elsevier, 1983.

Frederick Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, 1951.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Anand Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. *Advances in Neural Information Processing Systems*, 37:90132–90159, 2024.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development*, 63 (4/5):2–1, 2019.

Ritesh Noothigattu, Dominik Peters, and Ariel D Procaccia. Axioms for learning from pairwise comparisons. *Advances in Neural Information Processing Systems*, 33:17745–17754, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022.

John W Payne. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387, 1976.

John W Payne, James R Bettman, and Eric J Johnson. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, 14(3):534, 1988.

Emil Persson, Arvid Erlandsson, Paul Slovic, Daniel Västfjäll, and Gustav Tinghög. The prominence effect in health-care priority setting. *Judgment and Decision Making*, 17(6):1379–1391, 2022.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. Psychological forest: Predicting human behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Anuj K Shah and Daniel M Oppenheimer. Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, 134(2):207, 2008.

Özgür Şimşek and Marcus Buckmann. Learning from small samples: An analysis of simple decision heuristics. *Advances in neural information processing systems*, 28, 2015.

Walter Sinnott-Armstrong, Joshua August Skorburg, et al. How ai can aid bioethics. *Journal of Practical Ethics*, 9(1), 2021.

Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

LL Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

Amos Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

Amos Tversky, Shmuel Sattath, and Paul Slovic. Contingent weighting in judgment and choice. *Psychological review*, 95(3):371, 1988.

Amos Tversky, Paul Slovic, and Daniel Kahneman. The causes of preference reversal. *The American Economic Review*, pp. 204–217, 1990.

Atsushi Ueshima and Hiroki Takikawa. Discovering novel social preferences using simple artificial neural networks. *Collabra: Psychology*, 10(1), 2024.

Christopher Wiedeman, Ge Wang, and Uwe Kruger. Modeling of moral decisions with deep learning. *Visual Computing for Industry, Biomedicine, and Art*, 3(1):27, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Milan Zorman, Milojka Molan Štiglic, Peter Kokol, and Ivan Malčić. The limitations of decision trees and automatic learning in real world medical decision making. *Journal of medical systems*, 21(6):403–415, 1997.

A PROOF COMPENDIUM

Theorem 2.4 (Axiomatic Factoring Characterization)**.** *We now explore the consequences of our axioms on characterizing the response function $H(\cdot,\cdot)$ of a decision-maker.*
*1. **Atomic Model** Suppose axiom 1, and also that $\mathcal{X}$ is a countable domain. Then there exists an* atomic rule *$h_{inn} : \mathcal{X} \to \mathbb{R}$ and $h_{out} : \mathbb{R} \to [0,1]$ s.t. $H(\cdot,\cdot)$ can be factored as*

$$H(x_1, x_2) = h_{out}\left(h_{inn}(x_1) - h_{inn}(x_2)\right) \quad .$$

*2. $\sigma(\cdot)$**-Transitivity** Suppose axioms 2–3 hold for some* continuous *transitivity law $f(\cdot,\cdot)$. Then there exists a symmetric continuous random variable with full support and CDF $\sigma(\cdot)$ s.t.*

$$H(x_1, x_3) = \sigma\left(\sigma^{-1}(H(x_1, x_2)) + \sigma^{-1}(H(x_2, x_3))\right) \quad .$$

Moreover, there exists some atomic rule *$h_{inn} : \mathcal{X} \to \mathbb{R}$ such that $H(\cdot,\cdot)$ can be factored as*

$$H(x_1, x_2) = \sigma\left(h_{inn}(x_1) - h_{inn}(x_2)\right) \quad .$$

*3. **Unconditional Factor Model**: Suppose as in item 2, and also axiom 4 (NC). Then there exist* inner functions *$h^i_{inn} : \mathcal{X}_i \to \mathbb{R}$ for $i \in [d]$ such that the atomic rule $h_{inn}(\cdot)$ and decision rule $H(\cdot,\cdot)$ factor as*

$$h_{inn}(x) = \sum_{i=1}^{d} h^i_{inn}(x^{(i)}) \quad , \quad H(x_1, x_2) = h_{out}\left(\sum_{i=1}^{d} h^i_{inn}(x_1^{(i)}) - h^i_{inn}(x_2^{(i)})\right) \quad .$$

*4. **Conditional Factor Model**: Suppose as in item 2, and also axiom 5 (CIC). Then there exist* conditional inner functions *$h^{i,\cdot}_{inn} : \mathcal{X}_i \to \mathbb{R}$ for $i \in [d]$ and* condition features *(context) $\omega \subseteq [d]$, $\omega_i = \omega \setminus \{i\}$, s.t.*

$$H(x_1, x_2) = h_{out}\left(\sum_{i=1}^{d} h^{i, x_1^{\omega_i}}_{inn}(x_1^{(i)}) - h^{i, x_2^{\omega_i}}_{inn}(x_2^{(i)})\right) \quad .$$

Proof. The structure of this result is a bit convoluted, due to its presentation in four parts. We first show item 1 in isolation using a straightforward explicit construction. We then show that the conditions of item 2 imply *symmetry* of the transitivity law, and the assumed addition of *continuity* of $f(\cdot,\cdot)$ is sufficient to yield item 2.

From there, we address items 3 & 4. Note that because NC is a special case of CIC with condition set $\omega = \emptyset$, we show only item 4, as item 3 is a special case of item 4.

We first show item 1. For countable $\mathcal{X}$, it's trivially true that each $x \in \mathcal{X}$ can be encoded as a power of 2. Then, $h_{\text{inn}}(x_1) - h_{\text{inn}}(x_2)$ is then 0 iff $x_1 = x_2$, which by axiom 1, necessitates $h_{\text{out}}(0) = \frac{1}{2}$. Otherwise, $x_1 \neq x_2$, and we can recover x_1 and x_2 uniquely (the larger by rounding the (absolute) difference up to a the nearest power of 2, i.e., by taking $\log_2(|h_{\text{inn}}(x_i) - h_{\text{inn}}(x_j)|)$ the smaller via subtraction, and the sign indicating which is which). Therefore, using this construction of h_{inn}, for an arbitrary mapping between $\mathcal{X}$ and $\mathbb{Z}$, we can construct a corresponding h_{out} that reconstructs any possible decision rule $H(\cdot,\cdot)$ over $\mathcal{X}$.

We now show item 2.

We first show that these assumptions imply $f(u, v) = f(v, u)$ for all $u, v \in (0, 1)$, i.e., the transitivity law is symmetric. Assume WLOG $u \geq \frac{1}{2}$ and $v \geq \frac{1}{2}$, as by complementarity and WT, the remaining cases can be derived. Now, select an arbitrary x_0 in $\mathcal{X}$. For any $\varepsilon \in (0, \frac{1}{2})$, by *codomain span*, there exists an infinite sequence $x_1, x_2, \ldots$ such that $H(x_{i-1}, x_i) = \frac{1}{2} + \varepsilon$, and moreover, let k_1, k_2 be the smallest integers such that $H(x_0, x_{k_1}) \geq u$ and $H(x_0, x_{k_1}) \geq v$, respectively. Note that for any k, $H(x_0, x_k)$ is an increasing function, achieving $\lim_{k\to\infty} H(x_0, x_k) = 1$. Furthermore, by continuity, in the limit as $\varepsilon \to 0$, $H(x_0, x_{k_1})$ and $H(x_0, x_{k_2})$ converge to u and v, respectively. However, the grouping of this sequence into a segment of k_1 points approximating u and k_2 points approximating v was arbitrary, and the weak transitivity operator $f(\cdot,\cdot)$, by construction, must be associative. We may thus conclude that, due to continuity and the limit as $\varepsilon \to 0$, $f(u, v) = f(v, u)$.

Now observe that, by the continuous deep sets theorem (Theorem 7 of (Zaheer et al., 2017)), since $f(\cdot,\cdot)$ is symmetric, there exist continuous functions $\rho, \phi : \mathbb{R} \to \mathbb{R}$ such that

$$H(x_1, x_3) = f(H(x_1, x_2), H(x_2, x_3)) = \rho(\phi \circ H(x_1, x_2) + \phi \circ H(x_2, x_3)) \quad .$$

Assume WLOG that $\phi(\frac{1}{2}) = 0$ (as any shifting to ϕ can be absorbed by ρ in the composition).

Complementarity implies reflexivity, i.e., it holds that $H(x_1, x_1) = 1 - H(x_1, x_1) = \frac{1}{2}$. Consequently, we observe the f Now, again by reflexivity, and applying our derived weak transitive law, it holds

$$\begin{aligned} H(x_1, x_2) &= \rho(\phi \circ H(x_1, x_2) + \phi \circ H(x_2, x_2)) \\ &= \rho(\phi \circ H(x_1, x_2)) && \phi \circ H(x_2, x_2) = \phi(\frac{1}{2}) = 0 \ . \end{aligned}$$

Thus $\rho(\phi(u)) = u$ for all $u \in (0, 1)$.

Now, we argue that $\phi(u) + \phi(1-u) = 0$, which implies $\phi(u) = -\phi(1-u)$. Observe the following cancellation: $H(x_1, x_1) = f(H(x_1, x_2), H(x_2, x_1)) = \rho(\phi(u)+\phi(1-u)) = \rho(0) = \frac{1}{2}$. By continuity, and the property $\rho(\phi(u)) = u$, this implies that ϕ is either an increasing or a decreasing function. Assume now that, WLOG, ϕ (and consequently ρ) are both increasing functions, as output-negation of ϕ may be counterbalanced by input-negation of ρ.

To show that ρ and ϕ are true inverses, we require also that $\phi(\rho(v)) = v$ for all $v \in \mathbb{R}$. We show this by proving that $\rho(\cdot)$ is a strictly increasing continuous function on $\mathbb{R} \to (0, 1)$. Because ϕ is a strictly increasing function, and $\rho(\phi(1)) = 1$, clearly $\rho(\infty) = 1$ and ρ is a strictly increasing continuous function *on the domain* $(\phi(0), \phi(1))$. The only issue is that, if $\phi(1) = -\phi(0) < \infty$, then ρ is only *weakly increasing* (constant) over the rest of $\mathbb{R}$. Now, suppose BWOC that $\phi(1) = c$ for some $c < \infty$. By codomain span and increasingness of ρ, there exist x_1, x_2, x_3, x_∞, such that $H(x_1, x_2) \neq H(x_1, x_3)$, $H(x_i, x_\infty) = 1$ for each i, and the weak transitivity law is violated, as $H(x_1, x_2)$ and $H(x_1, x_3)$ *can not possibly* both be computed as $f(1, 1)$. NB this impossibility does not depend on the functional form we derive: Rather, there is a fundamental incompatibility between weak transitivity and prediction with certainty. We thus conclude that the image of $\phi(\cdot)$ is $\mathbb{R}$, hence ϕ and ρ are true inverses, and both strictly monotonically increasing continuous functions. Moreover, ρ takes the form of the CDF of a symmetric continuous random variable with support $\mathbb{R}$.

Henceforth, we have operated purely in terms of the predicted probabilities, i.e., $H(\cdot, \cdot)$, but the result requires us to draw conclusions in terms of some $h_{\text{inn}}(x)$. We now argue for the existence of such a decomposition. Essentially, we "eliminate the middle man," as

$$\begin{aligned} H(x_1, x_3) &= f(H(x_1, x_2), H(x_2, x_3)) && \text{Weak Transitivity} \\ &= \rho(\phi \circ H(x_1, x_2) + \phi \circ H(x_2, x_3)) && \text{See Above} \\ &= \rho(\phi \circ H(x_1, x_2) - \phi \circ H(x_3, x_2)) && \text{Complementarity} \\ &= \rho\left((h_{\text{inn}}(x_1) - h'_{\text{inn}}(x_2)) - (h_{\text{inn}}(x_3) - h'_{\text{inn}}(x_2))\right) && \text{See Below} \\ &= \rho\left(h_{\text{inn}}(x_1) - h_{\text{inn}}(x_3)\right) \ . \end{aligned}$$

The step marked SEE BELOW is rather subtle, but observe that it must hold BWOC for some functions $h_{\text{inn}}, h'_{\text{inn}} : \mathcal{X} \to \mathbb{R}$, as if it did not, then there would exist some x_2, x'_2 such that $H(x_1, x_3) = f(H(x_1, x_2), H(x_2, x_3)) \neq f(H(x_1, x'_2), H(x'_2, x_3))$, which violates weak transitivity. Essentially, due to invertibility, $\phi \circ H(x_1, x_3) = (h_{\text{inn}}(x_1) - h'_{\text{inn}}(x_2)) - (h_{\text{inn}}(x_3) - h'_{\text{inn}}(x_2))$ must be invariant under choice of x_2. In the subsequent step, h'_{inn} is eliminated, which effectively illustrates that $h_{\text{inn}} = h'_{\text{inn}}$. Finally, let $h_{\text{out}}(u) = \phi(u)$, and item 1 is complete.

We now show item 4 (noting again that item 3 is a special case).

We first show a technical lemma: we assume x''' can be obtained from x_1 by changing two independent features not in the condition set ω, each partial change resulting in x'_2 and x''_2 (as in the CIC axiom). Observe that

$$\begin{aligned} \sigma^{-1} \circ H(x_1, x'''_2) &= \sigma^{-1} \circ H(x_1, x'_2) + \sigma^{-1} \circ H(x'_2, x'''_2) + \sigma^{-1} \circ H(x_1, x''_2) + \sigma^{-1} \circ H(x''_2, x'''_2) - \sigma^{-1} \circ H(x_1, x''_2) && \sigma\text{-Transitivity} \\ &= \left(\sigma^{-1} \circ H(x_1, x'_2) + \sigma^{-1} \circ H(x_1, x''_2)\right) + \left(\sigma^{-1} \circ H(x'_2, x'''_2) + \sigma^{-1} \circ H(x''_2, x'''_2) + \sigma^{-1} \circ H(x'''_2, x_1)\right) && \text{Algebra} \\ &= \left(\sigma^{-1} \circ H(x_1, x'_2) + \sigma^{-1} \circ H(x_1, x''_2)\right) + \\ &\quad \tfrac{1}{2}\left(\sigma^{-1} \circ H(x'_2, x'''_2) + \sigma^{-1} \circ H(x''_2, x'''_2) + \sigma^{-1} \circ H(x'_2, x_1) + \sigma^{-1} \circ H(x''_2, x_1)\right) && \text{Algebra} \\ &= \tfrac{1}{2}\left(\sigma^{-1} \circ H(x_1, x'_2) + \sigma^{-1} \circ H(x_1, x''_2)\right) + \tfrac{1}{2}\left(\sigma^{-1} \circ H(x'_2, x'''_2) + \sigma^{-1} \circ H(x''_2, x'''_2)\right) && \text{Algebra} \\ &= \sigma^{-1} \circ H(x_1, x'_2) + \sigma^{-1} \circ H(x_1, x''_2) && \text{See Below} \end{aligned}$$

The final step applies the assumed CIC axiom: The only way this assumption can hold is if $\left(\sigma^{-1} \circ H(x_1, x'_2) + \sigma^{-1} \circ H(x_1, x''_2)\right) = \left(\sigma^{-1} \circ H(x'_2, x'''_2) + \sigma^{-1} \circ H(x''_2, x'''_2)\right)$. This is because of symmetry, if we negate the equation, flipping x_1 and x_2''', the same must hold, hence the conclusion of the technical lemma.

We now chain the above result to derive the desideratum, i.e., item 2. Given $x_0', x_1', x_2', x_3', \ldots, x_d'$, define the *transitive chain operator* $f(\cdot, \cdot, \cdots, \cdot)$ as

$$f(x_0', x_1', x_2', x_3', \ldots, x_d') = f(\cdots(f(f(x_0', x_1'), x_2'), \cdots), x_d') = h_{\text{out}}\left(\sum_{i=1}^{d} h_{\text{inn}}(x_{i-1}') - h_{\text{inn}}(x_i')\right) \quad ,$$

where the RHS applies the structure of σ-transitivity.

Now, take $x_\omega'^{i}$ to be x_2^i if $i \in \omega$, x_1^i otherwise. Define a sequence x_j' such that each x_j' takes the previous x_{j-1}' (starting with x_ω' for $j = 1$) and changes one feature $k \notin \omega$ from x_1^k to x_2^k, such that $x_{d-|\omega|}' = x_2$.

We then chain the result over all items not in the condition set ω to obtain

$$\begin{aligned}
H(x_1, x_2) &= h_{\text{out}}\left(h_{\text{inn}}(x_1) - h_{\text{inn}}(x_2)\right) && \textsc{By Assumption} \\
&= f(x_1, x_{2,\omega}, x_{2,(1)}, x_{2,(2)}, \ldots, x_{2,(d-|\omega|)}) && \textsc{Transitivity Chain} \\
&= h_{\text{out}}\left(h_{\text{inn}}^{\omega, x_1^\omega}(x_1) - h_{\text{inn}}^{\omega, x_2^\omega}(x_2) + \sum_{i \notin \omega} h_{\text{inn}}^{i,\omega_i}(x_1^i) - h_{\text{inn}}^{i,\omega_i}(x_2^i)\right) && \sigma\text{-}\textsc{Transitivity} \\
&= h_{\text{out}}\left(\sum_{i \notin \omega} h_{\text{inn}}^{i,\omega_i}(x_1^i) - h_{\text{inn}}^{i,\omega_i}(x_2^i)\right) && \textsc{See Below} \\
&= h_{\text{out}}\left(\sum_{i=1}^{d} h_{\text{inn}}^{i,\omega_i}(x_1^i) - h_{\text{inn}}^{i,\omega_i}(x_2^i)\right) \quad . && \textsc{Let} h_{\text{inn}}^{i,\omega_i}(x) = 0 \text{ for all } i \in \omega
\end{aligned}$$

Note that the function h_{inn} may change from step to step above, the conclusion is only that such a decomposition exists. To see the step marked SEE BELOW, observe that the terms that condition on ω are capable of additively representing any function over ω, so the explicit first term, i.e., that involving h_{inn}^{ω}, may be omitted. Finally, observe that for the special case of noninteractivity, we have $\omega = \emptyset$, thus this first term always cancels out, and the above telescoping decomposition consists of exactly d terms, one per feature. □

We now show thm. 2.6.

Theorem 2.6 (Axiomatic Models)**.** *Suppose axioms 1–4. The following then restrict $\mathcal{H}_{inn}$ or each $\mathcal{H}_{inn}^{(i)}$.*
1. Suppose σ-linearity. *Then $\mathcal{H}_{inn}^{(i)} = \{x \mapsto wx | w \in \mathbb{R}\}$, thus $\mathcal{H} = \{x_1, x_2 \mapsto \sigma(\boldsymbol{w} \cdot (x_1 - x_2)) \,|\, \boldsymbol{w} \in \mathbb{R}^d\}$.*
2. Suppose monotonicity. *Then $\mathcal{H}_{inn}^{(i)} = \{h : \mathbb{R} \to \mathbb{R} \,|\, x \leq y \implies h(x) \leq h(y)\}$.*
*3. **Multivariate Monotonic Models**: If we assume monotonicity but relax noninteractive compositionality, then $\mathcal{H} = \{x_1, x_2 \mapsto \sigma(h(x_1) - h(x_2)) \,|\, h : \mathcal{X} \to \mathbb{R}$ s.t. $\vec{x} \preceq \vec{y} \implies h(\vec{x}) \leq h(\vec{y})\}$.*
*4. **Conditional GAM Tree** If we relax NC to CIC, then $h_{inn}(\cdot)$ may be represented as a hybrid model of a* decision tree */ GAM model, starting with a tree over $\mathcal{X}_\omega$, where each leaf contains a* GAM *over $\mathcal{X}_{\setminus\omega}$. Formally, we have $\mathcal{H}_{inn} = \left\{\sum_{i=1}^{d-|\omega|}\left(T_i(x^\omega)\right)\left((x^{\setminus\omega})^{(i)}\right) \,\middle|\, T \in \mathbb{R}^{|\omega|} \to (\mathbb{R} \to \mathbb{R})^{d-|\omega|}\right\}$.*

Proof. We first show item 1. We then find that it is more straightforward to show item 3, and finally to conclude item 2 as a corollary.

We first show item 1. Recall that the linearity assumption requires that there exists some $\boldsymbol{w}$ such that for all x_1, x_2, it holds

$$\sigma^{-1} \circ H(x_1, x_2) = \boldsymbol{w} \cdot (x_1 - x_2) \quad .$$

Consequently, the set of all feasible decision rules is parameterized by $\boldsymbol{w}$, in particular, applying $\sigma(\cdot)$ to both sides of the above, it holds

$$\mathcal{H} = \left\{x_1, x_2 \mapsto \sigma(\boldsymbol{w} \cdot (x_1 - x_2)) \,\middle|\, \boldsymbol{w} \in \mathbb{R}^d\right\} \quad .$$

Finally, observe that this multivariate hypothesis class decomposes as

$$\mathcal{H} = \left\{ x_1, x_2 \mapsto \sigma\left(\sum_{i=1}^{d} h_{\text{inn}}^i(x_1^i) - h_{\text{inn}}^i(x_2^i) \right) \middle| h_{\text{inn}}^i(x) = \boldsymbol{w}_i x, \boldsymbol{w} \in \mathbb{R}^d \right\} \ .$$

From this, we conclude both the form of $\mathcal{H}$ and of each $\mathcal{H}_{\text{inn}}^{(i)}$.

NB σ of σ-linearity and σ-transitivity must be identical over the domain (up to isomorphism), as σ-linearity and σ-transitivity would otherwise be mutually incompatible.

We now show item 3.

Recall that we assumed σ-transitivity (thm. 2.4 item 2). It thus holds that

$$\mathcal{H} = \left\{ x^{(1)}, x^{(2)} \mapsto \sigma\left(\sum_{i=1}^{d} h_{\text{inn}}^i(x_1^{(i)}) - h_{\text{inn}}^i(x_2^{(i)}) \right) \middle| h_{\text{inn}}^i \in \mathcal{H}_{\text{inn}}^{(i)} \right\} \ .$$

Monotonicity requires that if $x_1 \preceq x_2$, then

$$H(x_1, x_2) \leq \frac{1}{2} \ .$$

In our factoring, this is equivalent to

$$h_{\text{inn}}(x_1) \leq h_{\text{inn}}(x_2)$$

Consequently, the set of all such functions satisfying monotonicity is

$$\mathcal{H} = \{ x_1, x_2 \mapsto \sigma(h(x_1) - h(x_2)) \,|\, h : \mathcal{X} \to \mathbb{R} \text{ s.t. } \vec{x} \preceq \vec{y} \implies h(\vec{x}) \leq h(\vec{y}) \} \ .$$

NB: this does not specify σ, but due to the assumed factoring, σ must be a symmetric continuous CDF with full support.

We now show item 2.

As we now restore NC, we now begin with the stronger factored form

$$\mathcal{H} = \left\{ x^{(1)}, x^{(2)} \mapsto \sigma\left(\sum_{i=1}^{d} h_{\text{inn}}^i(x_1^{(i)}) - h_{\text{inn}}^i(x_2^{(i)}) \right) \middle| h_{\text{inn}}^i \in \mathcal{H}_{\text{inn}}^{(i)} \right\} \ .$$

We then apply the same reasoning as in item 3, now applied to x_1, x_2 that differ in only dimension i, to conclude that the ith model factor obeys

$$\mathcal{H}_{\text{inn}}^{(i)} = \{ h : \mathbb{R} \to \mathbb{R} \,|\, x \leq y \implies h(x) \leq h(y) \} \ .$$

This concludes the result. □

We now show thm. 2.7.

Theorem 2.7 (Proportionality of Choice)**.** *Suppose a probabilistic decision rule over $n \geq 3$ items that obeys the* choice axiom*. Then $H(\cdot, \cdot)$ admits a σ-transitive factoring with logistic $\sigma(\cdot)$.*

Proof. Suppose items A, B, and $C \in \mathcal{X}$. Essentially, we use the telescoping product

$$\frac{\mathbb{P}(A \text{ best})}{\mathbb{P}(C \text{ best})} = \frac{\mathbb{P}(A \text{ best})}{\mathbb{P}(B \text{ best})} \frac{\mathbb{P}(B \text{ best})}{\mathbb{P}(C \text{ best})}$$

Substituting in the proportionality axiom:

$$\frac{1}{\sigma(\sigma^{-1}(\mathbb{P}(C > B)) + \sigma^{-1}(\mathbb{P}(B > A)))} - 1 = \left(\frac{1}{\mathbb{P}(B > A)} - 1 \right) \left(\frac{1}{\mathbb{P}(C > A)} - 1 \right)$$

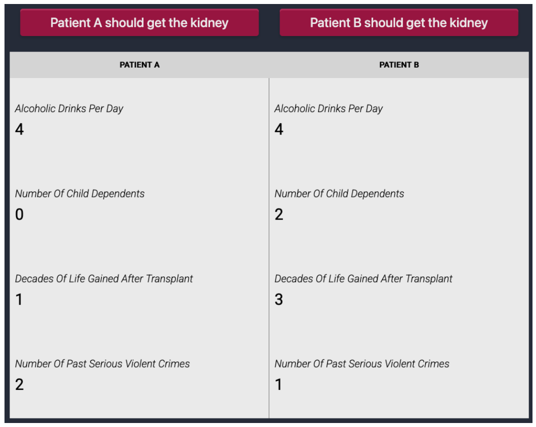

Figure 4: An example pairwise comparison presented to participants of Study One of the Kidney Allocation dataset.

Rearranging the equation:

$$
\begin{aligned}
\sigma(\sigma^{-1}(\mathbb{P}(C > B)) + \sigma^{-1}(\mathbb{P}(B > A))) &= \frac{1}{1 + \left(\frac{1}{\mathbb{P}(B>A)} - 1\right)\left(\frac{1}{\mathbb{P}(C>A)} - 1\right)} \\
&= \frac{1}{1 + \exp \circ \ln \left(\left(\frac{1}{\mathbb{P}(B>A)} - 1\right)\left(\frac{1}{\mathbb{P}(C>A)} - 1\right)\right)} \\
&= \text{logistic}\left(\ln\left(\frac{1}{\frac{1}{\mathbb{P}(B>A)} - 1}\right) + \ln\left(\frac{1}{\frac{1}{\mathbb{P}(C>A)} - 1}\right)\right) \\
&= \text{logistic}\left(\text{logit}\left(\mathbb{P}(B > A)\right) + \text{logit}\left(\mathbb{P}(C > A)\right)\right) \quad .
\end{aligned}
$$

We thus conclude that σ is the logistic function. $\square$

B Empirical Analysis – Other Details

B.1 Dataset and Preprocessing

Kidney Allocation Real-World Datasets. As mentioned earlier, the real-world dataset we used was collected by Boerstler et al. (2024) to test moral judgments *stability*. Across two studies, multiple participants were presented with several kidney allocation scenarios across 10 days (60 scenarios per day). In Study One, each kidney allocation scenario contained profiles of two kidney patients (Patient A and Patient B) described by four features: (a) number of dependents (0, 1, 2), (b) life decades gained from kidney transplant (1, 2, 3), (c) number of alcoholic drinks per day (0, 2, 4), and (d) number of crimes committed (0, 1, 2). In Study Two, each patient is described by five features: (a) number of elderly dependents (0, 1, 2), (b) life years gained from kidney transplant (1, 2, 3), (c) years waiting for the transplant (0, 2, 4), (d) weekly work hours post-transplant (0, 1, 2), and (e) obesity level (0, 1, 2, 3, 4). Participants were asked to decide who should get the kidney when only one was available and both patients were eligible. An example of one such pairwise comparison is presented in Figure 4.

Out of the 60 scenarios presented to each participant per day, several scenarios were repeated. This includes 6 scenarios that were repeated twice per day across all days to test for response stability and 2 scenarios that were used for attention checks. The remaining scenarios were randomly chosen. Since the repeated scenarios were non-random, we removed them from the dataset to ensure that the underlying data distribution for each participant corresponds to the uniform distribution over the input space. Sessions where participants failed attention checks were also removed from the dataset.

After the above steps, we had around 383 average responses from 15 participants in Study One and around 330 average responses from 40 participants in Study Two. This dataset is available under the CC-BY 4.0 license and is included with the Supplementary for the sake of completeness.

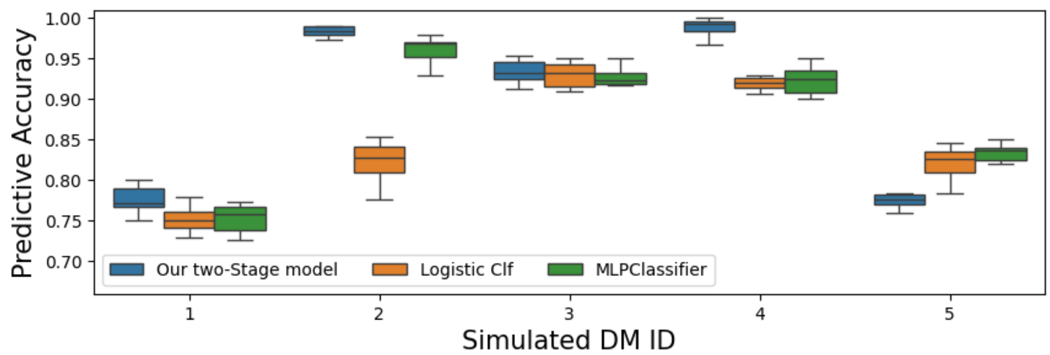

Figure 5: Comparison of our model vs baseline linear regression model for the synthetic dataset containing responses from five simulated decision-makers.

Synthetic Datasets. The synthetic dataset is constructed by simulating five decision-makers DM1–DM5, each using a different set of heuristics. The process for DM1 is described in Section 3. For DM2–DM5, the simulated decision processes are noted below. Recall that each decision-maker is presented with a pairwise comparison between two kidney patients, A and B, with features x_A and x_B. Each patient is described using the patient's number of dependents, life years gained from the transplant, years waiting for the transplant, and their past number of crimes.

DM2 simulation. For DM2, the decision process for any pairwise comparison between A and B can be described as follows: (a) choose the patient with non-zero dependents; (b) if both patients have non-zero dependents, then choose the patient with greater value for life years gained; (c) if equal, then choose the patient who has been waiting longer for the transplant; (d) if equal, then choose randomly. Hence, this decision-maker also uses thresholds on the number of dependents and employs other features mainly for tie-breaking.

DM3 simulation. For DM3, the decision process for any pairwise comparison between A and B can be described as follows: (a) transform life years gained feature to reflect *diminishing returns*, i.e. $z_A^{\text{life gained}} = \lfloor \log\left(x_A^{\text{life gained}}\right) \rfloor$ and $z_B^{\text{life gained}} = \lfloor \log\left(x_B^{\text{life gained}}\right) \rfloor$; (b) assign $z_A^{\text{life gained}}$ points to patient A and $z_B^{\text{life gained}}$ to patient B; (c) assign $x_A^{\text{dependents}}$ points to patient A and $x_B^{\text{dependents}}$ points to patient B; (d) assign one point each patient who was been waiting for 5 years or more; (e) sum up the points for each patient and choose one with greater number of points (ties broken randomly). The log-transformation used in this decision-making process captures the diminishing returns property associated with the life years gained feature. Additionally, this process also uses the tallying heuristic to essentially count the number of factors favoring each patient.

DM4 simulation. For DM4, the decision process for any pairwise comparison between A and B can be described as follows: (a) transform life years gained feature to reflect *diminishing returns*, i.e. $z_A^{\text{life gained}} = \lfloor \log\left(x_A^{\text{life gained}}\right) \rfloor$ and $z_B^{\text{life gained}} = \lfloor \log\left(x_B^{\text{life gained}}\right) \rfloor$; (b) choose patient with greater $z_{\cdot}^{\text{life gained}}$ value; (c) if equal, then choose the patient with more dependents; (d) if equal, then choose the patient who has been waiting longer for the transplant; (e) if equal, then choose randomly. This decision-maker also uses log-transformation for life years gained and other features for tie-breaking.

DM5 simulation. For DM5, the decision process for any pairwise comparison between A and B can be described as follows: (a) count how many features favor each patient and choose the patient favored by more features; (b) if equal, choose randomly. This decision-maker simply uses the tallying heuristic, choosing the option favored by more factors.

B.2 Implementation Details

For each decision-maker (real or simulated), we use a random 70-30 train-test split and report the summary statistics of predictive accuracy over test partitions across 20 repetitions.

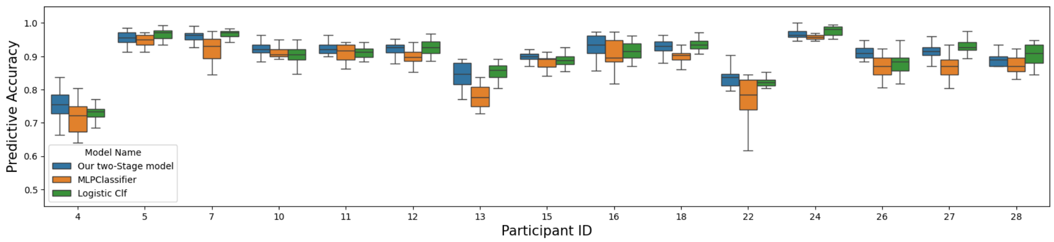

Figure 6: Comparison of our model vs baseline linear regression model for participants in Study One of the Kidney Allocation dataset.

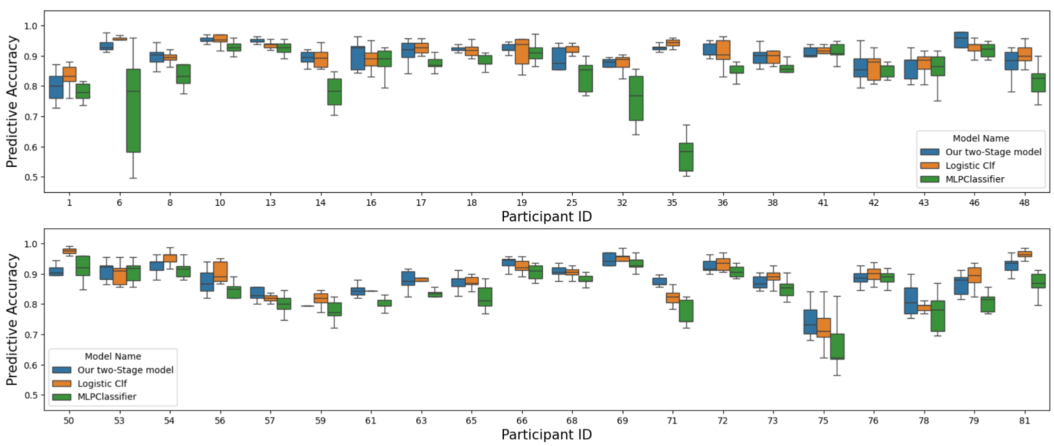

Figure 7: Comparison of our model vs baseline linear regression model for participants in Study Two of the Kidney Allocation dataset.

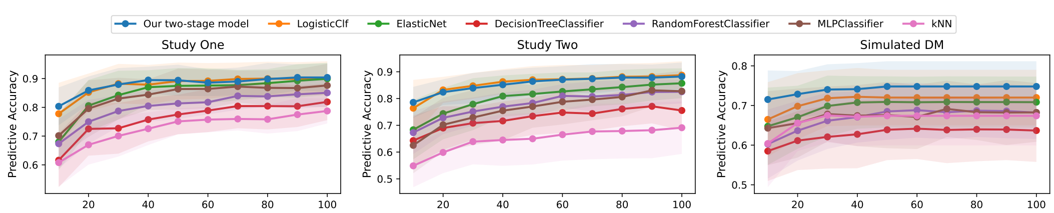

Figure 8: Comparison of our model vs baselines across increasing training size.

Our learning framework. For our framework, we learn editing functions $h^{i,\cdot}_{\text{inn}}$ for feature i that are designed to assign a score$\in \mathbb{R}$ to each value of feature i. As noted earlier, all $h^{\cdot,\cdot}_{\text{inn}}$ are constrained to be monotonic. Context ω is limited to one feature for the real-world dataset. The specific context is chosen using cross-validation. 20% of the training data is held out to use for validation purposes and the conditional two-stage model is learned for each setting of $\omega = \{i\}$ for all $i \in [d]$. The context $\omega = \{i^\star\}$ for which we achieve the smallest predictive error over the validation set is chosen as the final context for the corresponding participant. Context ω is kept empty for the synthetic dataset. For the dominance testing function $h_{\text{out}}(\cdot)$, we implement the simple tallying heuristic (i.e., the outputs from each feature-level editing function $h^{\cdot,\cdot}_{\text{inn}}$ are simply averaged). The final binary prediction is basically whether the $h_{\text{out}}(\cdot)$ is greater than 0 or not (with positive value favoring the patient on the left and negative value favoring the one on the right).

The two-stage model is trained by minimizing the chosen loss function (cross-entropy or hinge) with regularization and constraints. The regularization term quantifies the difference between editing functions learned for different values of the context variable ω. We use this regularizer to ensure that editing functions for any feature i corresponding to different context values are not too far from each other. As we noted earlier, ω is set to contain only one feature in our experiments on real-world datasets. For any two feature values $x^\omega = a$ and $x^\omega = b$, and for any other feature $i \notin \omega$, the regularization term measures $||h^{i,a}_{\text{inn}} - h^{i,b}_{\text{inn}}||$, for all a, b. The difference between the two editing functions is calculated by taking the squared norm of the difference between the outputs of these functions over all values in $\mathcal{X}_i$. Hence, the overall regularization term we use is

$$\lambda \cdot \sum_{i \notin \omega} \sum_{a,b} ||h^{i,a}_{\text{inn}} - h^{i,b}_{\text{inn}}||.$$

Here λ is the regularization parameter and is set to be 1e-3 in our experiments. Additionally, we impose monotonicity constraints on each of h_{inn} functions while optimizing the above loss; i.e., for all features i, $h^i_{\text{inn}}(a) \geq h^i_{\text{inn}}(b), \forall a > b \in \mathcal{X}_i$ OR $h^i_{\text{inn}}(a) \leq h^i_{\text{inn}}(b), \forall a > b \in \mathcal{X}_i$ (same constraints for context-based inner functions as well). We solve this constrained optimization problem using the Python SLSQP library (options 'ftol' and 'maxiter' are set to 1e-7 and 300, respectively).

Baseline details. The drift diffusion model was implemented using the Python PyDDM library with linear drift functions. The Bradley-Terry framework was implemented using the Python choix library, with a two-layer neural network scoring function.

We also implemented the following supervised modeling strategies to compare our method against. (a) Logistic Classifier – we implement the standard logistic classification approach, but impose a symmetry constraint by regressing over feature differences across the pairwise comparison; (b) Elastic-net Classifier – using the standard logistic classification approach over all given features with L1 and L2 regularization, with L1 ratio (scaling between L1 and L2 penalties) set to be 0.5. (c) Decision Tree Classifier – over all features with Gini splitting criterion; (d) Linear SVM – over all features with L2 penalty and squared hinge loss; (e) kNN – over all features with $n = 5$ neighbors; (f) Random Forest Classifier – over all features with 100 estimators; (g) MLP Classifier – over all features with two hidden layers of 10 nodes each.

Computing resources used. All experiments were run on a MacBook M2 system with a 16GB memory.

B.3 Additional Empirical Results

Performances and models for simulated decision-makers DM2–DM5. As mentioned earlier, we also created a synthetic dataset containing responses from five simulated decision-makers. The descriptions of the simulated DMs are provided above. In this section, we report additional results for these simulations.

First, the performance of our model, Logistic Classifier, and MLP Classifier are presented in Figure 5. Note for all but DM5, our model has better predictive accuracy than baselines.

Individual participant-level performances. Across all participants, the performance of our two-stage model, Logistic Classifier, and MLP Classifier are reported in Figure 6 for Study One and in Figure 7 for Study Two (other baselines excluded here for presentation clarity). The plots show that our model has comparable accuracy to the Logistic Classifier for all participants across the real-world study datasets.

Performance variation across training data sizes. We also assess the performance of each model across variations of training size from 10 to 100. The results are reported in Figure 8. Our model reaches high

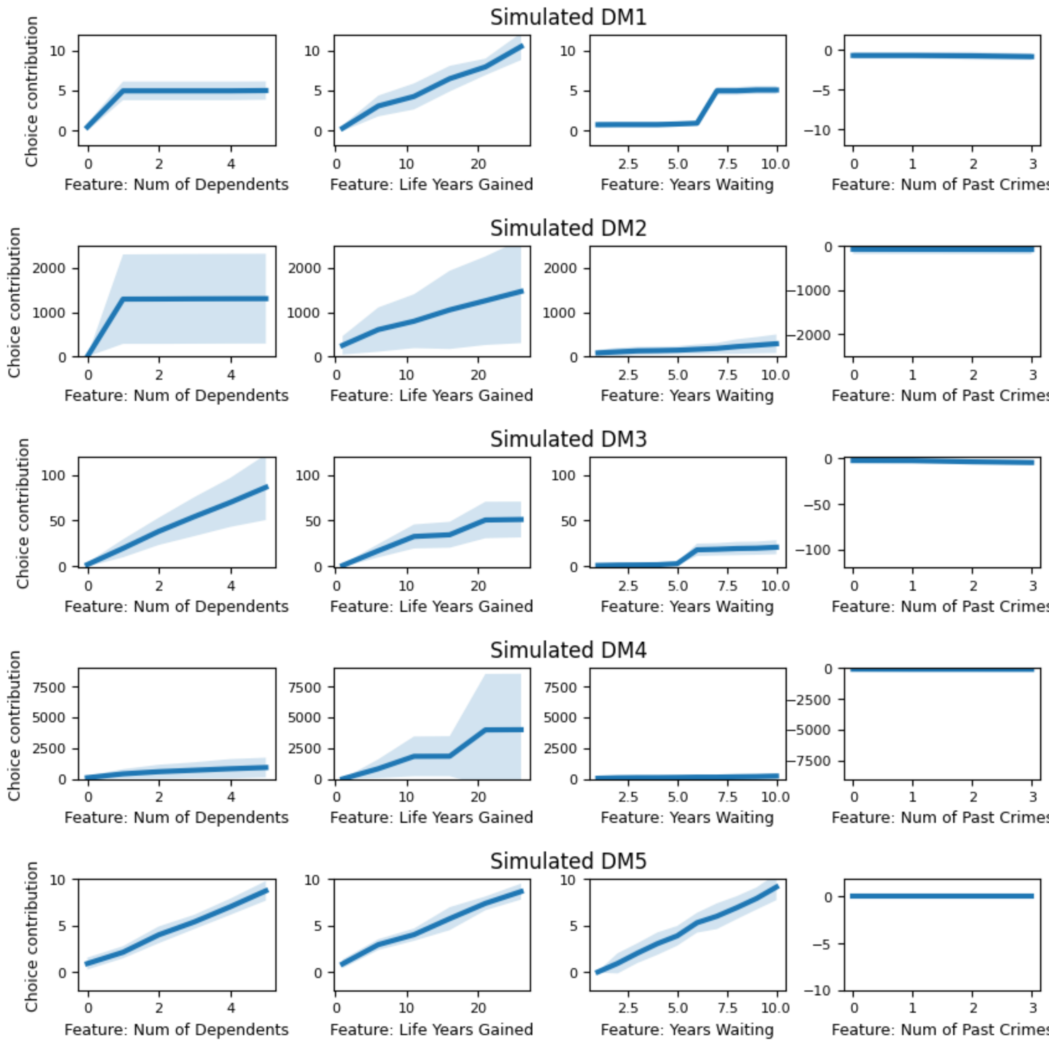

Figure 9: Editing rules learned by our approach for all simulated decision-makers DM1–DM5.

accuracy levels faster than other models, with significant improvement in training performance compared to baselines for the simulated decision-makers.