Aligning with Heterogeneous Preferences for Kidney Exchange

Rachel Freedman
University of California, Berkeley
rachel.freedman@berkeley.edu

Abstract

AI algorithms increasingly make decisions that impact entire groups of humans. Since humans tend to hold varying and even conflicting preferences, AI algorithms responsible for making decisions on behalf of such groups encounter the problem of *preference aggregation*: combining inconsistent and sometimes contradictory individual preferences into a representative aggregate. In this paper, we address this problem in a real-world public health context: *kidney exchange*. The algorithms that allocate kidneys from living donors to patients needing transplants in kidney exchange matching markets should prioritize patients in a way that aligns with the values of the community they serve, but allocation preferences vary widely across individuals. In this paper, we propose, implement and evaluate a methodology for prioritizing patients based on such *heterogeneous* moral preferences. Instead of selecting a single static set of patient weights, we learn a distribution over preference functions based on human subject responses to allocation dilemmas, then sample from this distribution to dynamically determine patient weights during matching. We find that this methodology increases the average rank of matched patients in the sampled preference ordering, indicating better satisfaction of group preferences. We hope that this work will suggest a roadmap for future automated moral decision making on behalf of heterogeneous groups.

1 Introduction

As AI algorithms become increasingly powerful and more widely deployed, it is vital that they act in a way that aligns with human values. Unfortunately, in most real-world domains, people do not unanimously agree on a single set of "human values" that AI algorithms can model and instantiate. Instead, groups of humans tend to hold varying and even conflicting moral preferences, and AI algorithms responsible for making decisions on behalf of these *heterogeneous* groups must aggregate and arbitrate between these preferences.

Many existing approaches to *preference aggregation* for AI rely on determining a single representative objective or decision for the AI to implement [9, 16, 6]. However, humans are known for their variable and contradictory preferences, meaning that many individuals will hold preferences that differ greatly from the mean. Better techniques are required to model such heterogeneous human preferences, implement them in AI algorithms, and measure their satisfaction in practice.

One domain in which this is particularly apparent is that of *kidney exchange*. In a kidney exchange, patients who need kidney transplants and have found willing but medically incompatible donors are matched and exchange kidneys with other such incompatible patient-donor pairs [12]. Many countries, including the United States [7], the United Kingdom [10] and much of Europe [4] use algorithms developed in the AI community to automate this matching. Since the prognosis for patients who do not receive kidney transplants is quite poor, these automated decisions have life-or-death consequences and great moral import. It is therefore vital that these allocation decisions are made in a way that aligns with societal values. Previous work has sought to learn a single static utility function that kidney allocation algorithms can use to prioritize certain types of matching [8]. However, that work disregards the empirical heterogeneity in human ethical judgements in this domain. We seek to instead model this heterogeneity by developing a methodology that represents the full distribution of human judgements.

In this work, we draw on preference aggregation and social welfare theory to design, implement and evaluate a methodology for autonomously allocating kidneys to patients in matching markets based on the heterogeneous moral
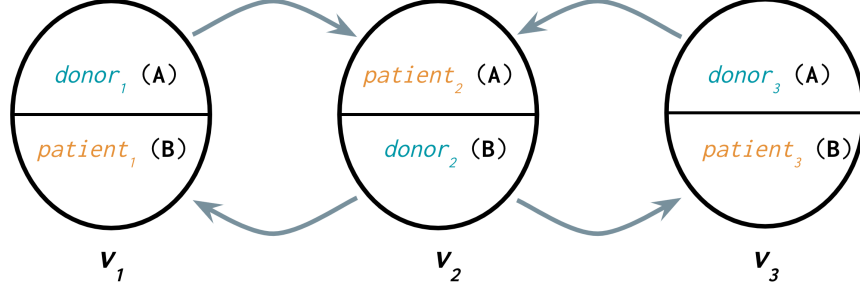
Figure 1: Example compatibility graph. Donor and patient blood types are in parentheses and arrows indicate possible valid donations. This graph has two valid donation cycles: $c_A = \{v_1, v_2\}$ and $c_B = \{v_2, v_3\}$. However, both contain v_2, so only one set of donations can take place.

preferences expressed by surveyed human participants. We propose an alternate model for aggregating preferences drawn from the economics literature and an alternate domain-specific measure of group welfare based on individual preference rankings. We show that our proposed model, which aggregates individual preferences into a distribution over utility functions instead of a single function, improves the average rank of the matched kidney donations in individuals' preference orderings, without reducing the number of patients that can be matched overall. Incorporating the model into this real-world AI system leads to more beneficial outcomes according to our proposed measure of social welfare.

We hope that this work will both highlight the preference aggregation challenges present in many allocative AI systems, and serve as a roadmap for developing systems that directly address these challenges in other real-world contexts.

2 Kidney Exchanges

2.1 Graph Formulation

In a *kidney exchange*, patients who need a kidney transplant and donors who are willing to donate to them but are medically incompatible can be matched with other such patient-donor pairs [12]. For example, if the donor of pair i is compatible with the patient of pair j, and the donor of pair j is likewise compatible with the patient of pair i, they can form a *matching*, in which donor d_i donates to patient p_j and donor d_j donates to patient p_i.

In the standard formulation, a kidney exchange is described by a *compatibility graph* $G = \langle V, E \rangle$ [12, 13]. We construct one vertex v for each patient-donor pair, then add a directed edge $e_{i,j}$ from v_i to v_j if d_i is compatible with p_j. A cycle c is a possible sequence of valid transplants, in which each donor in the cycle donates a kidney and each patient receives one. A matching M is a set of disjoint cycles. An example compatibility graph is shown in Figure 1. Each oval in the figure represents a vertex, and each arrow represents a directed edge signifying donor-patient compatibility. There are two cycles in this particular compatibility graph: $c_A = \{v_1, v_2\}$ and $c_B = \{v_2, v_3\}$. However, these two cycles are not disjoint, because they share vertex v_2. The v_2 donor cannot donate both of their kidneys, so these exchanges cannot both take place. This compatibility graph therefore has two valid matchings, $M_A = \{c_A\}$ and $M_B = \{c_B\}$, each with cardinality 2.

2.2 Clearing Algorithm

The *clearing house problem* in kidney exchange is to find the optimal valid matching, according to some utility function [1]. Finding valid matchings with a finite limit on cycle lengths is NP-hard [1] and difficult to approximate [3], so this problem is typically solved by formulating it as a linear program (LP) and solving it with an LP-solver such as CPLEX.

We typically assign a weight w_e to each edge e to represent the utility of that particular donation taking place. In the national US exchange, these weights are set ad-hoc by a committee [15], but in this work we will adapt an alternative method that learns these weights based on human responses to allocation dilemmas [8]. The clearing house problem is to find the optimal matching M^* which maximizes some utility function $u : M \to \mathbb{R}$. This is typically formalized as the graph-theoretic problem of finding the maximum weighted cycle cover $u(M) = \sum_{c \in M} \sum_{e \in c} w_e$. To solve this

Profile	Age	Drinking	Cancer
1	30	rare	healthy
2	30	frequently	healthy
3	30	rare	cancer
4	30	frequently	cancer
5	70	rare	healthy
6	70	frequently	healthy
7	70	rare	cancer
8	70	frequently	cancer

Table 1: Patient profile descriptions enumerated and (arbitrarily) numbered by Freedman *et al.*

via linear programming, let $C(L)$ be the set of all cycles of length no more than L, let $w_c = \sum_{c \in C(L)} w_e$, create an activation variable $x_c \in 0, 1$ for each cycle $c \in C(L)$, then solve the following linear program:

$$\max \sum_{c \in C(L)} w_c\, x_c \qquad s.t. \qquad \sum_{c:v \in c} x_c \leq 1 \quad \forall v \in V. \tag{1}$$

using an LP-solver such as CPLEX. The final matching M is the set of cycles c with an activation $x_c = 1$. If all edge weights w_e are equal, then solving this LP gives a maximum-cardinality matching. In cases where there are multiple valid matchings of equivalent cardinality, such as M_A and M_B in Figure 1, this LP-solver must choose between them randomly. However, if the edge weights are set according to some utility function, then the solution can prioritize certain types of matches. This allows us to incorporate societal preferences.

2.3 Incorporating Preferences

Previous work has attempted to improve the matching prioritization in kidney exchanges based on sampled human ethical preferences [8]. All else equal, it is obviously morally preferable to save lives by matching as many patient-donor pairs as possible. However, in cases such as the one in Figure 1, there can be multiple maximum-cardinality matchings. In this case, the algorithm requires a utility function that distinguishes between them, ideally in a way that aligns with human values. The US national kidney exchange attempts to do this, but they prioritize matches in an opaque and ad-hoc fashion via committee [15]. This excludes most of the societal members who will actually participate in the exchange from the discussion, and leaves the committee with the still-unsolved problem of designing a utility function that captures the ethical preferences of an entire society. Freedman *et al.* propose an alternative methodology for *learning* domain-relevant ethical preferences from actual human decisions in kidney allocation dilemmas, revising the LP in Eq 1 to take these into account, and then evaluating the impact on a simulated exchange. Our work proposes an improvement on Freedman *et al.*'s methodology for aggregating preferences and evaluating results, so we will briefly outline their full methodology here.

Freedman *et al.* conducted two surveys on participants from Amazon's Mechanical Turk platform ("MTurk"). The first survey asked participants ($N = 100$) to read a brief description of the kidney transplant waiting list process, and then asked them to propose which patient characteristics they thought "morally ought" to be used to prioritize patients. The three most popular categories of responses were "age", "health – behavioral" (including aspects of health believed to be under personal control, such as diet and drinking), and "health – general" (including aspects of health unrelated to kidney disease). The second survey asked a new set of participants ($N = 289$) to decide how to allocate kidneys between pairs of fictional patient "profiles" that vary according to these attributes. In order to make the profiles more concrete, drinking behavior ("1 alcoholic drink per month" or "5 alcoholic drinks per day") was used as a proxy for behavioral health, and cancer ("skin cancer in remission" or "no other major health problems") was used as a proxy for general health. For example, a sample question asked participants to choose between "Patient W.A. [who] is 30 years old, had 1 alcoholic drink per month (prior to diagnosis), and has no other major health problems" and "Patient R.F. [who] is 70 years old, had 5 alcoholic drinks per day (prior to diagnosis), and has skin cancer in remission". They

defined 8 such patient profiles, with characteristics described in Table 1, and asked each participant to compare each pair of profiles. We will use these profile descriptions and this preference data in our own work.

Freedman *et al.* used the *Bradley-Terry Model* ("BT Model") to estimate a "BT-score" for each patient profile. The BT model assumes that each profile x has an underlying score p_x that represents the value that survey participants collectively place on donating to a patient with that profile. Under this model, the probability that profile i will be preferred to profile j is:

$$P(i > j) = \frac{p_i}{p_i + p_j} \tag{2}$$

Patient profiles that are almost always selected by our survey participants (such as profile 1 in Table 1) will therefore have the highest scores, and profiles that are rarely selected (such as profile 8), will therefore have the lowest scores. Freedman *et al.* use this model to estimate a single set of scores based on the pooled pairwise comparisons from every survey respondent. This allows them to aggregate all preferences into a single set of scores.

They then revised the LP from Eq 1, setting the weight of each edge $e_{i,j}$ to be the BT-score of the recipient p_j and adding a cardinality constraint to require that the LP still only produce maximum-cardinality matchings. Let $w_{BT(v)}$ be the BT-score of the patient in vertex v, and let Q be the maximum matching cardinality possible for the compatibility graph. Then the revised LP is:

$$
\begin{aligned}
\max \quad & \sum_{c \in C(L)} \left[\sum_{(u,v) \in c} w_{BT(v)} \right] x_c \\
\text{s.t.} \quad & \sum_{c:v \in c} x_c \leq 1 \qquad \forall v \in V \\
& \sum_{c \in C(L)} |c| x_c \geq Q
\end{aligned}
\tag{3}
$$

They evaluated this revised algorithm on a simulated kidney exchange and found that it matched significantly more of the higher-scoring profiles and significantly fewer of the lower-scoring ones. However, this methodology relies on the assumption that societal preferences are sufficiently homogeneous to be captured by a single static utility function. An algorithm using this methodology will always choose to save a patient of profile 1 over a patient of profile 2. However, the preferences expressed in the survey data actually varied greatly, and participants did sometimes prefer patients of profile 2 to profile 1. Presumably the preferences of a representative sample of the actual US population would be even more heterogeneous. In this sense, both the static profile scoring and the assessment of the algorithm by the proportion of each profile matched are flawed. In this work, we improve upon both of these elements by removing the requirement for a single utility function and developing an alternate methodology for modifying and evaluating the algorithm.

3 Incorporating Heterogeneous Preferences

In our work we improve upon the methodology presented in Section 2.3 by removing the unrealistic assumption that societal preferences can be captured by a single utility function. We propose 1) an alternative preference aggregation method that better captures the variation in expressed preferences, 2) modifications to the kidney allocation algorithm to take this new preference aggregation into account, and 3) an alternative evaluation metric for the resulting matchings that lends more consideration to individual welfare.

3.1 Preference Aggregation Model: BLP

Instead of learning a single score for each profile as in previous work [8], we use the *Berry-Levinsohn-Pakes Model* ("BLP Model") to estimate a *distribution* over possible utility functions. We propose that learning and sampling from this distribution better satisfies individual preferences than learning a single utility function. The BLP model is an extension of the logit discrete choice model that is widely used in estimating consumer discrete-choice demand for differentiated products [2, 11]. When we apply this model to kidney exchange, the "consumers" are members of the population that the exchange serves, and the "products" are the patients who may potentially be matched with donors. Using this model allows us to predict how the general population wants the exchange to prioritize patients.

For a graph $G = \langle V, E \rangle$, we wish to define a utility function $\mathcal{U} : V \to \mathbb{R}$ that determines the utility of the patient in each vertex receiving a utility function. The BLP model defines the utility function $\mathcal{U}(v) = X_{p(v)}^\top \beta + \epsilon$ where $X_{p(v)} = \{\mathbb{1}(v_{age} = 30), \mathbb{1}(v_{drinking} = rare), \mathbb{1}(v_{cancer} = healthy)\}$ are the binary features of the patient profile of vertex v, $\beta \sim \mathcal{N}(\mu, \Sigma)$ gives the weight of each feature, and ϵ is an error term following a type-II extreme value distribution.

We use maximum likelihood estimation to fit the distribution parameters (μ, Σ) to the pairwise comparison survey data gathered by Freedman *et al.* Let P be the set of all patient profiles described in Table 1, and for each pair of profiles $i, j \in P$, let $c_k(i,j)$ be survey respondent k's preferred profile. This allows us to define the likelihood function

$$L_k(\mu, \Sigma \mid c_k) = \mathbb{E}_{\beta \sim \mathcal{N}(\mu, \Sigma)} \left[\prod_{i,j} \frac{\exp(X_{c_k(i,j)}^\top \beta)}{\exp(X_i^\top \beta) + \exp(X_j^\top \beta)} \right] \tag{4}$$

and to estimate the maximum likelihood distribution parameters

$$\hat{\mu}, \hat{\Sigma} = \underset{\mu, \Sigma}{argmax} \frac{1}{N} \sum_{k=1}^{N} \log(L_k(\mu, \Sigma \mid c_k)) \tag{5}$$

3.2 Algorithm

Each time a new patient-donor pair enters the exchange, we add a corresponding vertex u to the graph, randomly sample a $\beta_u \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ from the learned distribution, and weight outgoing edges $u \rightarrow v$ using the resulting "BLP function": $BLP(u,v) = X_{p(v)}^\top \beta_u + \epsilon_{uv}$. In this way, we represent the full distribution of preferences. Note that the BLP function indicates a random sample from the surveyed population's preference distribution – it does not represent the preferences of donor u specifically. Letting $w_{BLP(u,v)}$ be the score that vertex u's sampled BLP function places on donating to the patient in vertex v, we modify the LP in Eq 3 to be:

$$\begin{array}{ll} \max & \sum_{c \in C(L)} \left[\sum_{(u,v) \in c} w_{BLP(u,v)} \right] x_c \\ \text{s.t.} & \sum_{c:v \in c} x_c \leq 1 \qquad \qquad \forall v \in V \\ & \sum_{c \in C(L)} |c| x_c \geq Q \end{array} \tag{6}$$

3.3 Evaluation Metric

We further define the $rank$ of a donation $u \rightarrow v$ to be the position of v's patient profile in the preference ordering induced by u's BLP function. For example, if the BLP function associated with vertex u weights the profile of the patient in vertex v above all other profiles, $rank(u,v) = 1$. Conversely, if the BLP function weights the profile below the other seven possible patient profiles, $rank(u,v) = 8$. In this context, *rank* functions as a proxy for individual welfare because it represents the extent to which an individual's domain-relevant values were fulfilled. We claim that the average rank of matched donations is a better measure of the extent to which an algorithm values individual welfare than the proportion of each profile matched because the ranks of all matches depend on the full BLP distribution. In contrast, the proportion matched measure relies on the false assumption that everyone's preferences are better satisfied if patients with higher BT-scores are matched more often.

We run both algorithms on a simulated kidney exchange, along with a third algorithm that weights all donations equally as a baseline. We evaluate the resulting matchings both on the proportion of each profile matched, and on our proposed average rank measure. We find that our proposed algorithm consistently outperforms both others on the rank measure, suggesting that it better represents the full distribution of societal preferences.

4 Experiments

4.1 Conditions

We tested three versions of the matching algorithm: the baseline one that weights all donations equally, the one with a single utility function described in Section 2.3, and one with a distribution over utility functions proposed in Section 3.1.

Condition 1: EQUAL The EQUAL algorithm matches kidney exchange participants using the LP in Eq 1. That is, it weights all participants equally and chooses randomly amongst the highest-cardinality matchings. We use this condition as a baseline because it describes the case in which ethical preferences are not incorporated into the algorithm at all.

Condition 2: `HOMOGENEOUS`   The `HOMOGENEOUS` algorithm matches participants using the LP in Eq 3. It assigns edge weights based on the BT-score of the recipient, relying on the assumption that individual preferences are sufficiently homogeneous to be captured by a single static utility function. This is the algorithm proposed by Freedman *et al.*. See Table 2 for the weights used in the `EQUAL` and `HOMOGENEOUS` conditions.

Condition 3: `HETEROGENEOUS`   The `HETEROGENEOUS` algorithm matches participants using the LP in Eq 6. It samples a BLP function when each vertex is added to the graph, normalizes the scores produced by that function to the range $[0, 1]$, and uses that function and the profile of the recipient to weight each new outgoing edge. This allows for the possibility that heterogeneous individual preferences are better captured by a distribution than by a single utility function. This is the novel algorithm that we propose in this work.

Profile ID	EQUAL	HOMOGENEOUS
1	1.000	1.000
2	1.000	0.103
3	1.000	0.236
4	1.000	0.036
5	1.000	0.070
6	1.000	0.012
7	1.000	0.024
8	1.000	0.003

Table 2: Patient profile weights for the `EQUAL` and `HOMOGENEOUS` experimental conditions. The `EQUAL` algorithm values all profiles equally, so all have weight 1.0. However, the `HOMOGENEOUS` algorithm weights profiles according to their BT-scores. The `HETEROGENEOUS` algorithm samples BLP functions throughout matching and so does not have a static weight for each profile.

4.2   Measures

We evaluate each algorithm according to both the measure we propose in Section 3 and the measure used by Freedman *et al.*

Average Rank   The *average rank* of a matching is the average rank of each donation in the matching, where $rank(u, v)$ of a donation $u \rightarrow v$ is as defined in Section 3.3. Recall that lower ranks indicate higher preference satisfaction and, since there are 8 profiles, all possible ranks fall in the range $[1.0, 8.0]$. For each run of each algorithm, we recorded the average rank of every matching in the simulation, then averaged these to get the average rank value for that algorithm.

Proportion Matched   The *proportion matched* of a profile is the proportion of patients of that type that entered the kidney exchange pool and were subsequently matched. A proportion matched of 100% means that all patients of that type were matched, and a proportion matched of 0% means that none were. For each run of each algorithm, we recorded the number of patients with each profile that entered the pool and the number of patients of each profile that were eventually matched, and used this to calculate proportion matched.

4.3   Experimental Setup

We built a simulator[1] to mimic daily matching using the `EQUAL`, `HOMOGENEOUS`, or `HETEROGENEOUS` algorithms based on previously developed tools [7, 8]. Each simulated "day", some pairs enter the pool, some pairs exit the pool, and then the matching algorithm is run on the pairs that remain. The unmatched pairs remain in the pool to potentially be matched in the future. For each algorithm, we executed 50 runs of 5 years of simulated daily matching, and recorded the average matching rank and profile proportions matched for each run.

---

[1]All code for this paper can be found in the `Variation` package of `github.com/RachelFreedman/KidneyExchange`
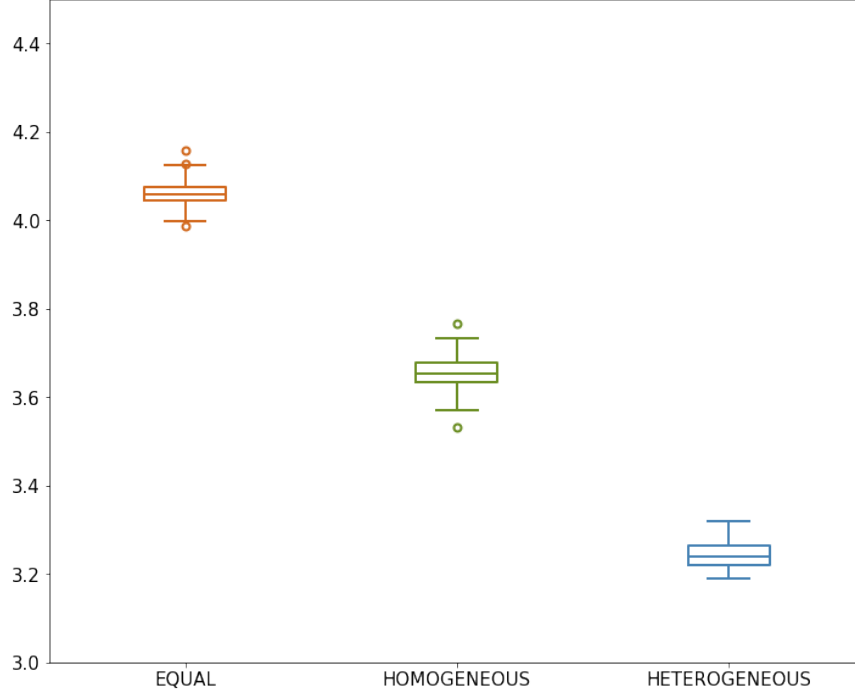
Figure 2: Average rank of donations in each simulated run ($N = 50$). Lower ranks indicate more-preferred matches. The `HETEROGENEOUS` algorithm produces the lowest average ranks (median = 3.24), followed by the `HOMOGENEOUS` algorithm (median = 3.66), then the `EQUAL` algorithm (median = 4.06).

5   Results

Average Rank   Since lower donation ranks indicate that the recipient is higher in the sampled preference ordering, we propose that algorithms that induce lower average ranks better satisfy population preferences. As expected, the proposed `HETEROGENEOUS` algorithm consistently produces matchings with the lowest average rank (Figure 2). The `HOMOGENEOUS` algorithm produces the next-lowest average rank, followed by the `EQUAL` algorithm, which produces the highest average rank. This is because the `EQUAL` algorithm weights all edges *equally*, matching recipients without any consideration of the personal characteristics used to define their weight. The `HOMOGENEOUS` algorithm improves upon this by considering the characteristics of donation recipients, but fails to approximate preferences as closely as `HETEROGENEOUS`.

Proportion Matched   Freedman *et al.* quantified the impact of their modified algorithm by comparing the proportions of patients of each profile type matched by their algorithm against the proportions matched by the unmodified algorithm, so for the sake of comparison we do the same. The proportions of each profile matched by the `EQUAL` algorithm, the `HOMOGENEOUS` algorithm (proposed by Freedman *et al.*) and the `HETEROGENEOUS` algorithm (proposed in this work) are shown in Figure 3.

Since it doesn't take patient profiles into account, the `EQUAL` algorithm matched approximately the same percentage of patients across all profiles. Since it prioritizes patients solely based on profile, the `HOMOGENEOUS` algorithm matched the more popular profiles (1-3) more often and the less popular profiles (4-8) less. Notably, the `HOMOGENEOUS` algorithm almost always matches patients of profile 1, indicating that a patient's profile can be one of the major factors in determining whether they receive a kidney. However, the `HETEROGENEOUS` algorithm prioritizes patients not directly based on their profile, but based on the sampled BLP function's *valuation* of their profile. As a result, this algorithm still tends to match more of the commonly-preferred profiles and fewer of the commonly-dispreferred ones, but sometimes samples a BLP function from the tails of the distribution that prioritizes patient profiles very differently.

If the survey preferences had been completely homogeneous, then the `HETEROGENEOUS` algorithm would have produced the same results as the `HOMOGENEOUS` algorithm. However, because preferences expressed in the survey data sometimes differ from the utility function used for the `HOMOGENEOUS` algorithm, sometimes different matches are made. For example, while most survey participants preferred patient profile 1 to all other profiles, some did not, so the
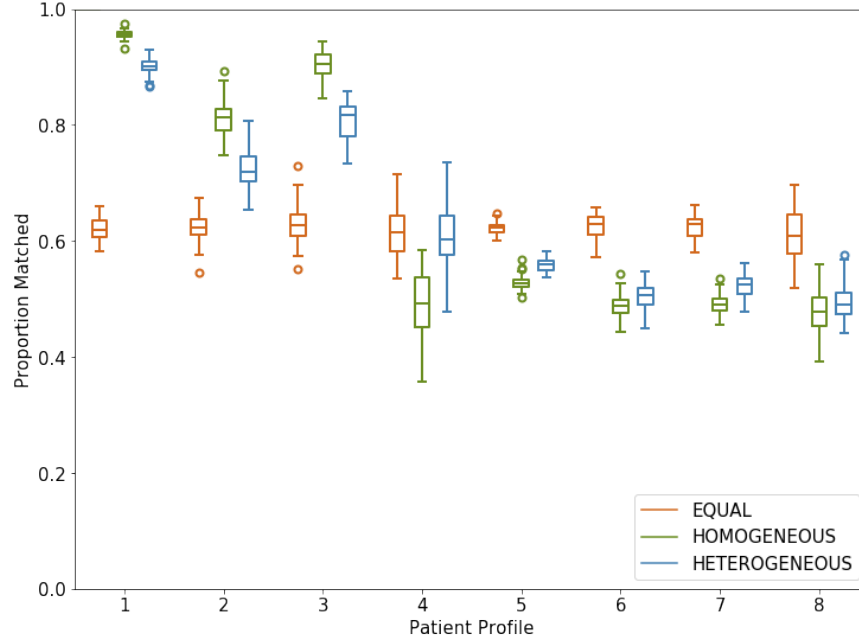
Figure 3: Proportion of patients of each profile matched in each simulated run ($N = 50$). All algorithms match approximately 62% of patients overall, but the `HOMOGENEOUS` algorithm disproportionately matches profiles with higher BT-scores, the `EQUAL` algorithm matches all profiles approximately equally, and the proportions matched by the `HETEROGENEOUS` algorithm lie between the other two.

`HETEROGENEOUS` algorithm respects this heterogeneity by sometimes prioritizing matching other profiles over profile 1. This difference in matching is a further indication that our proposed algorithm more faithfully represents the full distribution over preferences.

6 Discussion

Faithfully instantiating the collective values of groups with heterogeneous individual preferences is a frequent challenge for real-world AI systems. For example, we commonly use AI systems to allocate scarce resources – such as kidney donors, food donations [9] and interview slots [14] – amongst group members in a way that we hope maximizes group welfare. Moreover, our roads may soon be populated with autonomous vehicles, which will have to make moral tradeoffs – such determining who to sacrifice in unavoidable collisions [5] – based on the complex and often contradictory moral frameworks of the communities in which they operate. It is therefore vital that the AI community develop techniques for faithfully aggregating such heterogeneous preferences and use them to develop socially beneficial AI systems.

In this paper, we proposed, instantiated, and evaluated one such technique for incorporating heterogeneous ethical preferences into a specific real-world AI system: an algorithm for matching patient-donor pairs in kidney exchange. Instead of weighting all potential kidney recipients equally, deciding how to prioritize them in an opaque and ad-hoc way [15], or prioritizing them based on a single static utility function [8], we proposed learning a distribution over surveyed preferences and then sampling from this distribution for dynamic weighting during matching. We furthermore proposed donation rank as a better measure of preference satisfaction. We implemented our proposed algorithm and compare it to predecessor algorithms on a kidney exchange simulation, finding that our algorithm better satisfies survey participant preferences.

Our model was estimated based on preference data elicited from MTurk survey participants, who are assuredly not representative of society in general. Future work should elicit preferences from a more representative sample, and perhaps privilege preferences expressed by domain experts and stakeholders such as doctors, policy-makers and kidney exchange participants. However, we believe that our sample was not *more* heterogeneous than the US population as a whole, so we expect the challenge of heterogeneity and our methodology to continue to be relevant for this expanded sample. Moreover, since even a representative sample of the general public would still lack relevant domain-specific

knowledge about kidney transplants, future work should also investigate methodologies that allow domain experts to correct or moderate the outcomes of this process.

We hope that the challenges highlighted and methodology prototyped in this work will suggest a roadmap for developing techniques for automating moral decision making in other domains.

Acknowledgements

References

[1] David J Abraham, Avrim Blum, and Tuomas Sandholm. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 295–304, 2007.

[2] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[3] Peter Biro and Katarina Cechlarova. Inapproximability of the kidney exchange problem. *Information Processing Letters*, 101(5):199–202, 2007.

[4] Péter Biró, Lisa Burnapp, Bernadette Haase, Aline Hemke, Rachel Johnson, Joris van de Klundert, and David Manlove. Kidney exchange practices in Europe, 2017. First Handbook of the COST Action CA15210: European Network for Collaboration on Kidney Exchange Programmes.

[5] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[6] Vince Conitzer, Markus Brill, and Rupert Freeman. Crowdsourcing societal tradeoffs. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, pages 1213–1217, 2015.

[7] John P Dickerson and Tuomas Sandholm. Futurematch: Combining human value judgments and machine learning to match in dynamic environments. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[8] 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*, page 103261, 2020.

[9] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Ritesh Noothigattu, Daniel See, Siheon Lee, Christos-Alexandros Psomas, et al. Webuildai: Participatory framework for fair and efficient algorithmic governance. *Preprint*, 2018.

[10] David Manlove and Gregg O'Malley. Paired and altruistic kidney donation in the UK: Algorithms and experimentation. *ACM Journal of Experimental Algorithmics*, 19(1), 2015.

[11] Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

[12] Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney exchange. *The Quarterly journal of economics*, 119(2):457–488, 2004.

[13] Alvin E Roth, Tayfun Sönmez, et al. A kidney exchange clearinghouse in new england. *American Economic Review*, 95(2):376–380, 2005.

[14] Candice Schumann, Zhi Lang, Jeffrey Foster, and John Dickerson. Making the cut: A bandit-based approach to tiered interviewing. In *Advances in Neural Information Processing Systems*, pages 4641–4651, 2019.

[15] UNOS. Revising kidney paired donation pilot program priority points, 2015. OPTN/UNOS Public Comment Proposal.

[16] Hanrui Zhang and Vincent Conitzer. A pac framework for aggregating agents' judgments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2237–2244, 2019.