The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Adapting a Kidney Exchange Algorithm to Align with Human Values

Rachel Freedman
Duke University
rachel.freedman@duke.edu

Jana Schaich Borg
Duke University
js524@duke.edu

Walter Sinnott-Armstrong
Duke University
ws66@duke.edu

John P. Dickerson
University of Maryland
john@cs.umd.edu

Vincent Conitzer
Duke University
conitzer@cs.duke.edu

Abstract

The efficient allocation of limited resources is a classical problem in economics and computer science. In kidney exchanges, a central market maker allocates living kidney donors to patients in need of an organ. Patients and donors in kidney exchanges are prioritized using ad-hoc weights decided on by committee and then fed into an allocation algorithm that determines who get what—and who does not. In this paper, we provide an end-to-end methodology for estimating weights of individual participant profiles in a kidney exchange. We first elicit from human subjects a list of patient attributes they consider acceptable for the purpose of prioritizing patients (e.g., medical characteristics, lifestyle choices, and so on). Then, we ask subjects comparison queries between patient profiles and estimate weights in a principled way from their responses. We show how to use these weights in kidney exchange market clearing algorithms. We then evaluate the impact of the weights in simulations and find that the precise numerical values of the weights we computed matter little, other than the ordering of profiles that they imply. However, compared to not prioritizing patients at all, there is a significant effect, with certain classes of patients being (de)prioritized based on the human-elicited value judgments.

Introduction

As AI is deployed increasingly broadly, AI researchers are increasingly confronted with moral implications of their work. The pursuit of simple objectives, such as minimizing error rates, often results in systems that have unintended consequences when they confront the real world, such as discriminating against certain groups of people (O'Neil 2017). It would be helpful for AI researchers and practitioners to have a general set of principles with which to approach these problems (Wallach and Allen 2008; Tolchinsky et al. 2012; Greene et al. 2016; Conitzer et al. 2017; Noothigattu et al. 2018).

One may ask why any moral decisions should be left to computers at all. There are multiple possible reasons. One is that the decision needs to be made so quickly that calling in a human for the decision is not feasible, as would be the case for a self-driving car having to make a split-second decision about whom to hit (Bonnefon, Shariff, and Rahwan 2016). Another reason could be that each individual decision by itself is too insignificant to bother a human, even though all the decisions combined may be highly significant morally—for example, if we were to consider the moral impact of each advertisement shown online. A third reason is that the moral decision is hard to decouple from a computational problem that apparently exceeds human capabilities. This is the case in many machine learning applications (e.g., should this person be released on bail?), but also in other optimization problems.

We are interested in one such problem: the clearing house problem in *kidney exchanges*. In a kidney exchange, patients who need a kidney transplant and have a willing but incompatible live donor may attempt to trade their donors' kidneys (Roth, Sonmez, and Unver 2004). Once these people appear at an exchange, we face a highly complex problem of deciding who matches with whom. In some exchanges, this matching problem is solved using algorithms developed in the AI community: the United States (Dickerson and Sandholm 2015), the United Kingdom (Manlove and O'Malley 2015), the Netherlands (Glorie, van de Klundert, and Wagelmans 2014), and so on (Biró et al. 2017).

In this paper, we investigate the following issue. Suppose, in principle, that we prioritize certain patients over others—for example, younger patients over older patients. To do so clearly would be a morally laden decision. How should this affect the role of the AI researcher developing these systems? From a pure algorithmic perspective, it may seem that there is little more to this than to change some weights in the objective function accordingly. But we argue that our job, as AI researchers, does not end with this simple observation. Rather, we should be closely involved with the process for determining these weights, both because we can contribute technical insights that are useful for this process itself, and because it is our responsibility to understand the consequences to which these weights will lead.

Our Contributions

In this paper, we provide an end-to-end methodology for estimating weights of individual patient profiles in a kidney exchange, where these weights are used only for tiebreaking purposes (i.e., when multiple solutions give the maximal number of transplants). We execute this methodology in a limited fashion as a proof of concept, and evaluate the re-

sults in simulations. (Executing our methodology in such a way that we would advocate directly adopting the results in practice would require substantially more effort and participation from other parties, as will become clear.)

We first elicit from human subjects a list of patient attributes they consider acceptable for the purpose of prioritizing patients in kidney exchanges (e.g., most subjects did not find race an acceptable attribute for prioritization). Then, we ask subjects comparison queries between patient profiles that differ only on acceptable attributes, and estimate weights from their responses. We show how to use these weights in kidney exchange market clearing algorithms, to break ties among multiple maximum-sized solutions. We then evaluate the impact of the weights in simulations. We find that the precise numerical values of the weights we computed matter little, other than the ordering of profiles that they imply. However, compared to not prioritizing patients at all, there is a significant effect. Specifically, the difference is experienced by donor-patient pairs that have an "underdemanded" (Ashlagi and Roth 2014; Toulis and Parkes 2015) combination of blood types; for them, their chances rise or drop significantly depending on their tiebreaking weights.

Kidney Exchange Model

We briefly review the standard mathematical model for kidney exchange and techniques from the AI community used to clear real kidney exchanges, and then give illustrative examples where tiebreaking would or would not play a role.

Graph Formulation

In this work, as is standard (Roth, Sonmez, and Unver 2004; Roth, Sönmez, and Ünver 2005a; 2005b), we encode an instance of a kidney exchange as a directed *compatibility graph* $G = (V, E)$. We first construct one vertex for each patient-donor pair in the pool. Then, we construct an edge e from vertex v_i to vertex v_j if the patient in v_j wants and is compatible with the donor kidney of v_i. A paired donor is willing to give her kidney if and only if the patient in her vertex v_i receives a kidney.

Most fielded exchanges also assign a weight w_e to an edge e; the function determining the weight for an edge is often opaque and set in an ad-hoc fashion by a committee;[1] it roughly represents the utility to v_j of obtaining v_i's donor kidney, but can also be used to (de)prioritize specific classes of patient or donor, as we discuss later. A cycle c represents a possible sequence of transplants, with each vertex in c obtaining the kidney of the previous vertex. We use the term *k-cycle* to refer to a cycle with exactly k pairs. For example, the compatibility graph in Figure 1 includes two possible 2-cycles: a 2-cycle between vertex v_1 and v_2, and a different 2-cycle between vertex v_2 and v_3. In kidney exchange, cycles of length at most some small constant L (typically, $L \in \{2, 3, 4\}$) are allowed—all transplants in a cycle must be performed simultaneously so that no donor backs out after his patient has received a kidney but before he has donated his kidney.

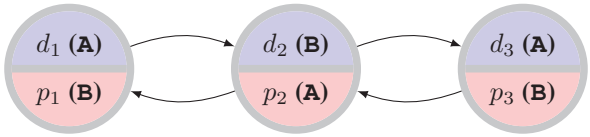

Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parantheses.

Many fielded kidney exchanges gain great utility through the use of *chains* (Montgomery et al. 2006; Rees et al. 2009; Anderson et al. 2015; Ashlagi et al. 2017). Chains start with an altruist donor donating her kidney to a patient, whose paired donor donates his kidney to another patient, and so on. In the standard model, altruistic donors are represented in the same way as patient-donor pairs, but with so-called "dummy" patients who are compatible with every patient-donor pair, yet do not require a kidney. In this way, altruists and patient-donor pairs—as well as cycles and chains—can be treated similarly in optimization models.

A *matching* M is a set of disjoint cycles and chains in the compatibility graph G. There can be length limits on these cycles and chains, as discussed above, resulting in a smaller set of *legal matchings*. The cycles and chains must be disjoint because no donor can give more than one of her kidneys (some recent work explores multi-donor donation (Ergin, Sönmez, and Ünver 2017; Farina, Dickerson, and Sandholm 2017) but we do not consider this here). Given the set of all legal matchings $\mathcal{M}$, the *clearing house problem* is to find a matching M^* that maximizes utility function $u : \mathcal{M} \to \mathbb{R}$. Formally:

$$M^* \in \arg\max_{M \in \mathcal{M}} u(M)$$

Kidney exchanges typically use a *utilitarian* utility function that finds the maximum weighted cycle cover (i.e., $u(M) = \sum_{c \in M} \sum_{e \in c} w_e$). This can favor certain classes of patient-donor pairs while marginalizing others, a behavior we investigate later in this paper in the context of setting specific edge weights. Alternate utility functions can be used to enforce incentive properties via mechanism design (Ashlagi and Roth 2014; Li et al. 2014; Hajaj et al. 2015; Blum et al. 2017; Mattei, Saffidine, and Walsh 2017).

Clearing Kidney Exchanges

We briefly discuss optimization methods for clearing kidney exchanges; later, we show how to augment these methods to incorporate the ideas in this paper. The standard clearing house problem for finite cycle cap $L > 2$ (even without chains) is NP-hard (Abraham, Blum, and Sandholm 2007; Biró, Manlove, and Rizzi 2009), and is also hard to approximate (Biró and Cechlárová 2007; Luo et al. 2016; Jia et al. 2017). Thus, fielded kidney exchanges use integer program (IP) formulations to solve this difficult combinatorial optimization problem.

[1] For a look into the inner workings of the process that sets edge weights, we direct the reader to a recent report by the UNOS US-wide kidney exchange (UNOS 2015).

The first approach to clearing large kidney exchanges, due to Abraham, Blum, and Sandholm (2007), built a custom branch and price (Barnhart et al. 1998) integer program solver; generalizations of, and improvements on, their basic model have addressed scalability issues (Dickerson, Procaccia, and Sandholm 2013; Glorie, van de Klundert, and Wagelmans 2014; Anderson et al. 2015; Dickerson et al. 2016). We build a similar model in this work.

Formally, denote the set of all chains of length at most K and cycles of length no greater than L by $C(L, K)$. Create a binary variable $x_c \in \{0, 1\}$ for every $c \in C(L, K)$, and let $w_c = \sum_{e \in c} w_e$; then, solve the following integer program:

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

The final matching is the set of chains and cycles c such that $x_c = 1$. In this paper, we compare to a baseline where all edge weights are 1, so that a maximum-cardinality solution is sought. We then break ties in these solutions based on prioritization weights determined according to the procedure outlined in this paper.

Tiebreaking and Prioritization: Examples

Consider again the compatibility graph given in Figure 1. Here, there is one pair with a patient of blood type A and a donor of blood type B, and two pairs with a patient of blood type B and a donor of blood type A. One of the latter two pairs will have to remain unmatched; either way, we obtain a solution of maximum cardinality (two vertices matched). The standard algorithm may choose either solution; which one is chosen depends on details of the solver. We may wish to break the tie based on other attributes of the two patients with blood type B, such as their age. We will explore this in this paper.

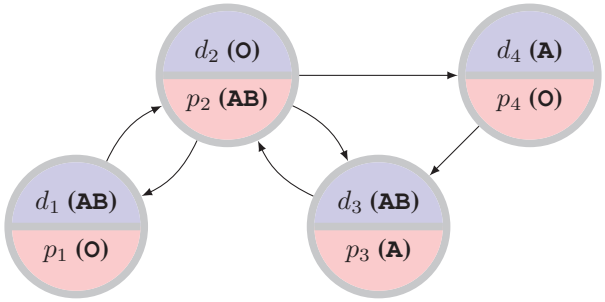

Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

Now, consider the graph in Figure 2. This graph has two maximal solutions (a solution is maximal if it is not possible to include any other vertices without dropping others from the solution). One consists of the 3-cycle with vertices AB-O, O-A, and A-AB (patient listed first in each case). The other consists of the 2-cycle with vertices AB-O and O-AB. The standard algorithm must choose the 3-cycle, because it matches more vertices. While in principle one might consider choosing the 2-cycle, arguing that (due to other attributes) it is more important to save the patient from the O-AB vertex than it is to save *both* the patient from the O-A vertex *and* the patient from the A-AB vertex, in this paper we will not do so; we will always choose the 3-cycle, no matter what the values of the additional attributes are.

Determining and Using Prioritization Weights

In this section, we describe our procedure for computing prioritization weights and integrating them into the algorithm for clearing kidney exchanges.

Selecting Attributes

First, we determined which patient attributes to include in our model by assessing which attributes a pool of human participants found acceptable to use for this purpose. The attributes were generated by the participants in an open-ended survey to minimize experimenter bias. Specifically, participants (N = 100) were recruited through the online platform Amazon Mechanical Turk, and asked to read a brief description of the kidney transplant waiting list process. Each participant then reported which attributes they thought should, and should not, be used to prioritize kidney transplant patients. Each participant received $0.85 compensation for their participation. Participants' responses were sorted into attribute categories by two independent coders. Attributes that the algorithm already takes into account, such as patient-donor medical compatibility, were discarded. The number of participants who mentioned each of the remaining attributes were counted.

The three attribute categories that the most participants thought should be used to prioritize patients were "Age", "Health - Behavioral" (aspects of health that are generally perceived to be controllable, such as diet and drug use), and "Health - General" (aspects of health that are generally perceived to be involuntary and are unrelated to kidney disease, such as cancer prognosis). There was a sharp drop-off in popularity between the third most popular category, "Health – General" (reported 44 times) and the fourth most popular one, "Dependents" (whether the patient had dependents, reported 18 times), so only the first three attribute categories were selected for inclusion in the next stage of the study.

Evaluating Pairwise Comparisons

We next gathered data on how people use the three top participant-generated attributes to prioritize patients. We administered a "Kidney Allocation Survey" to a new cohort of participants recruited through MTurk. In this survey, we turned each of the three chosen attributes into a binary one, as described in Table 1 below. The Age alternatives represent an adult nearer to the beginning of their adult life (but still of legal drinking age, 30 years old) or nearer to the end (70 years old). For a health-behavioral attribute, we chose alcohol consumption as a (potentially) controllable behavior that contributes to kidney disease. The indicated amount of alcohol consumption is specified to occur "prior to diagnosis," because drinking afterward disqualifies patients from the waiting list. Skin cancer was chosen as the "unhealthy"

alternative for the Health-General characteristic because it is a specific, well-known disease that may or may not be fatal.

Attribute	Alternative 0	Alternative 1
Age	30 years old (**Y**oung)	70 years old (**O**ld)
Health - Behavioral	1 alcoholic drink per month (**R**are)	5 alcoholic drinks per day (**F**requent)
Health - General	no other major health problems (**H**ealthy)	skin cancer in remission (**C**ancer)

Table 1: The two alternatives selected for each attribute. The alternative in each pair that we expected to be preferable was labeled "0", and the other was labeled "1".

Because there are three binary attributes, there are eight possible patient profiles. These eight unique patient profiles were enumerated and assigned ID numbers. In the survey, participants were asked to choose between pairs of these profiles. Participants (N = 289) were again recruited through MTurk. They read a short description of how kidney waiting lists work, and were asked to imagine that they were responsible for allocating a single kidney to one of two fictional patients. Each participant was then presented with all $\binom{8}{2} = 28$ possible pairs of profiles, in random order, and asked in each case to select the patient that they believed should receive the kidney. For half of the participants, the profile with the smaller ID number appeared on the screen above the profile with the larger ID number for each question ("original order"), and for the other half of the participants this order was reversed ("reversed order"), to counteract possible ordering or screen location effects. Each participant received $1.00 compensation for participating in this part of the study.

Summary of Responses Aggregate responses to the Kidney Allocation Survey are summarized below. The "Preferred" column reports the percentage of times that each profile was chosen in all the comparisons in which it appeared.

Profile	Age	Drinking	Cancer	Preferred
1 (YRH)	30	rare	healthy	94.0%
3 (YRC)	30	rare	cancer	76.8%
2 (YFH)	30	frequently	healthy	63.2%
5 (ORH)	70	rare	healthy	56.1%
4 (YFC)	30	frequently	cancer	43.5%
7 (ORC)	70	rare	cancer	36.3%
6 (OFH)	70	frequently	healthy	23.6%
8 (OFC)	70	frequently	cancer	6.4%

Table 2: Profile ranking according to Kidney Allocation Survey responses. The "Preferred" column describes the percentage of time the indicated profile was chosen among all the times it appeared in a comparison.

As expected, there was a clear preference for profile 1 (30 years old, 1 alcoholic drink per month, no other major health problems), and a clear preference against profile 8 (70 years old, 5 alcoholic drinks per day, skin cancer in remission). The preference for profile 3 (skin cancer in remission but minimal drinking) over profile 2 (healthy other than heavy drinking), and similarly 7 over 6, suggests that participants put greater weight on the health-behavioral attribute than on the health-general one. (Of course, this observation may not generalize to other health-behavioral and health-general attributes, such as drinking soda and paralysis.)

Estimating Profile Scores

We performed statistical modeling of participants' pairwise comparisons between patient profiles in order to obtain weights for each profile. We used the Bradley-Terry model, which treats each pairwise comparison as a contest between a pair of players (Bradley 1984). Under this model, each player i has a score p_i, representing its skill or value. Given two players i and j with respective scores p_i and p_j, the probability that player i will win the contest is:

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

In our context, each player is a patient profile, and each contest is a human subject comparison between two profiles. In each "contest", the profile that a participant selects is the one that wins. For example, suppose there are only two profiles, 1 and 2; in comparisons between them, one subject selected 1 and the next subject selected 2. For profile scores p_1 and p_2, the probability of this would be $p_1 p_2/(p_1 + p_2)^2$, which is maximized when $p_1 = p_2$. The BT scores (that we estimate based on our data) then constitute one measure of the value that the survey participants collectively place on "saving" each profile. The higher this value, the more likely a randomly selected participant is to select that profile over another. We can then use these scores as weights. (One may wonder whether perhaps it would be better to somehow transform—e.g., take the square root of—the weights first; one of our experiments below suggests this would make almost no difference.) This estimation procedure constitutes a specific way to *aggregate* the human subjects' moral judgments into a single weight for each profile; the strategy of using social choice theory to aggregate moral preferences for decision making has already been proposed by several groups (Greene et al. 2016; Conitzer et al. 2017; Noothigattu et al. 2018), and our specific approach fits well in the literature on interpreting voting as a method for statistically estimating an underlying truth (for an overview, see Elkind and Slinko (2015)).

We estimate BT scores in two different ways. One is to estimate scores directly for all profiles, so one profile's score is not constrained by the scores of other profiles. The second is to consider the importance of the individual attributes and let the score of profile i be a linear function of these:

$$\sum_{r=1}^{p} \beta_r x_{ir} + U_i$$

where x_{ir} is profile i's value for attribute r, and we estimate the β_r (importance of attribute r). The U_i are individual error terms where $U_i \sim N(0, \sigma^2)$, resulting in correlation between comparisons that share a common profile.

We used the `BTm()` function in the BradleyTerry2 package in R to estimate profile scores $p_1, \ldots, p_8$ based on the 8092 pairwise comparisons, both directly and as a function of the estimated scores of their three attribute values. The most-preferred profile, profile 1 in both cases, was assigned a score of 1. The results are in the following table.

Profile	Direct	Attribute-based
1 (YRH)	1.000000000	1.00000000
3 (YRC)	0.236280167	0.13183083
2 (YFH)	0.103243396	0.29106507
5 (ORH)	0.070045054	0.03837135
4 (YFC)	0.035722844	0.08900390
7 (ORC)	0.024072427	0.01173346
6 (OFH)	0.011349772	0.02590593
8 (OFC)	0.002769801	0.00341520

Table 3: The patient profile scores estimated using the Bradley-Terry Model. The "Direct" scores correspond to allowing a separate parameter for each profile (we use these in our simulations below), and the "Attribute-based" scores are based on the attributes via the linear model.

Adapt Algorithm

The final step was to incorporate the obtained weights into the kidney exchange market clearing algorithm. Because our human subjects data and analysis do not involve comparisons between differing quantities of patient profiles (e.g., choosing two patients with profile 1 over three patients with profile 2), we feel it is inappropriate to use the weights for such decisions. We only use the weights to break ties between solutions of maximum cardinality.

To find a matching, our adapted (prioritized) algorithm first runs the basic IP-based algorithm due to Abraham, Blum, and Sandholm (2007) with unit edge weights (i.e., $w_e = 1 \; \forall e \in E$). Our algorithm records the number of patients that receive a kidney in this solution as Q, and adds a new constraint to the IP requiring that the solution includes at least Q vertices. We then re-solve the IP with a new objective, using the weights corresponding to the patient profile scores derived from the survey responses. Formally, with $|c|$ denoting the *number* of vertices in cycle c, $type : V \rightarrow \{1, \ldots, 8\}$ mapping a vertex to its patient's profile, and w_θ denoting the score of profile θ, we solve:

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

This results in a set of kidney exchange cycles that includes the maximum possible number of patients, but prioritizes patient profiles that the surveyed population preferred.

Experiments

Having described how we obtained weights and how we integrated these weights into the IP-based algorithm, we now describe our experiments testing the effects of our prioritizing algorithm in simulations.

Experimental Setup

Based on previously developed tools (Dickerson and Sandholm 2015), we built a simulator to mimic daily matching in a real-world kidney exchange pool.[2] In the simulation, each day, some incompatible patient-donor pairs enter the simulated pool and some depart. Then, a matching algorithm is run to match a subset of compatible patient-donor pairs. The remaining incompatible pairs stay in the pool for consideration on the next day (and possibly beyond). Finally, the matches formed the previous day are executed with a certain success probability, and the matched pairs are removed from the pool. The demographics of our simulated pool were designed to reflect the UNOS kidney exchange pool where possible, and otherwise the general US population.

Experiment 1: Matchings with pair scores

Experiment In the first experiment, we compared the patient-donor pairs (vertices) matched by the original algorithm, which treats all profiles equally and breaks ties arbitrarily, to the pairs matched by the "prioritized" algorithm, which breaks ties towards pairs with higher (patient) profile scores. We ran 20 simulations of daily matching over the course of 5 simulated years using both algorithms.

We hypothesized that the original algorithm would match pairs in approximately the same proportion for every profile, but that the prioritizing algorithm would match pairs with higher profile scores more often than pairs with lower scores. Moreover, we hypothesized that the pairs with the highest profile scores (profiles 1, 3, and 2) would be matched more often by the prioritizing algorithm than by the original algorithm, and that the pairs with the lowest profile scores (profiles 7, 6, and 8) would be matched more often by the original algorithm than the prioritizing algorithm.

Results The proportions of pairs of each profile type matched by the original and prioritizing algorithms are plotted in Figure 1 below. "Proportion Matched" is the proportion of pairs that entered the pool that were subsequently matched. Both algorithms matched approximately 61.7% of pairs overall. (This result does not follow immediately from the fact that both algorithms match the maximum number of pairs in each round, because which specific profiles are matched in a round will affect which profiles appear in future rounds, and consequently may affect how many can be matched in future rounds.)

The results support both of our hypotheses. First, the original algorithm, called "STANDARD" in Figure 3, matched pairs approximately 62% of the time, regardless of their profile, while the prioritizing algorithm, called "PRIORITIZED" in Figure 3, matched the pairs with profile 1, who had the highest profile scores, nearly twice as often as it matched pairs with profile 8, who had the lowest profile scores. Secondly, pairs with profiles 1, 3, and 2 were indeed matched substantially more often by the prioritizing algorithm than by the original algorithm, while pairs with profiles 7, 6, and 8 were indeed matched substantially less often

[2] All code for this paper can be found in the `Ethics` package of `github.com/JohnDickerson/KidneyExchange`.

by the prioritizing algorithm than by the original algorithm. Thus, the scores assigned by the prioritizing algorithm do have a substantial effect on which profiles get matched.

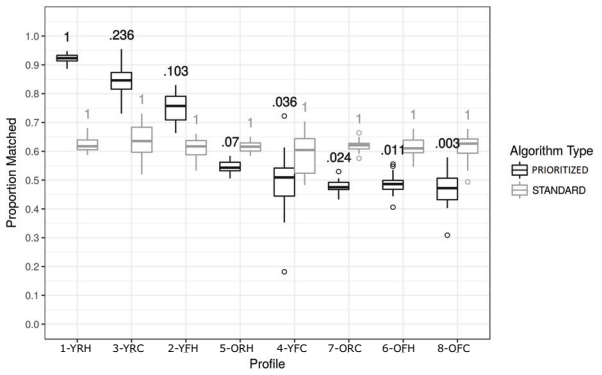

Figure 3: The proportions of pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally, it assigns each profile a score of 1. In this figure and later figures, each box represents the interquartile range (middle 50%), with the inner line denoting the median. The whiskers extend to the furthest data points within 1.5 $\times$ the interquartile range of the median, and the small circles denote outliers beyond this range.

Experiment 2: Matchings evaluated by blood type

Experiment To further explore how the modified algorithm prioritizes pairs with high profile scores at the expense of pairs with lower profile scores, we again ran 20 simulations of 5 simulated years of daily matching, this time recording the patient and donor blood types of each pair in addition to their profiles. We partitioned pairs into four established blood type classes motivated by large market analysis (Ashlagi and Roth 2014; Toulis and Parkes 2015). *Underdemanded* pairs were those that contain a patient with blood type O, a donor with blood type AB, or both, making them the most difficult to match. *Overdemanded* pairs contain a patient with blood type AB, a donor with blood type O, or both; *self-demanded* pairs contain a patient and donor with the same blood type; and *reciprocally demanded* pairs contain one person with blood type A, and one person with blood type B. These three classes are substantially easier to match. We hypothesized that the prioritizing algorithm primarily impacts underdemanded pairs, prioritizing underdemanded pairs with higher profile scores at the expense of underdemanded pairs with lower profile scores, while matching pairs that belong to the three other blood type classes at roughly the same high rates that the original algorithm does. The reasoning was that, intuitively, there is generally a scarcity of matching opportunities for the underdemanded pairs, but this is not so for the other types of pairs.

Results The results confirm our hypotheses. The proportions of underdemanded pairs matched are plotted in Figure 4. We found the proportions of overdemanded, self-demanded, and reciprocally demanded profiles matched to be fairly similar, so we grouped them together in Figure 5. The prioritizing algorithm matched underdemanded pairs with high profile scores substantially more often and underdemanded pairs with low scores substantially less often than the original algorithm did, but both algorithms matched pairs of other classes at roughly equal rates. This suggests that the primary difference between the algorithms lies in how they treat underdemanded pairs.

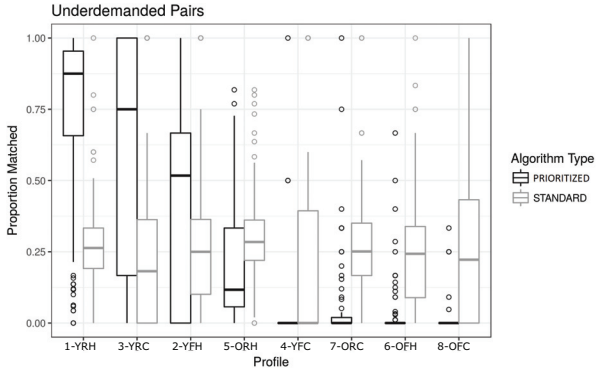

Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

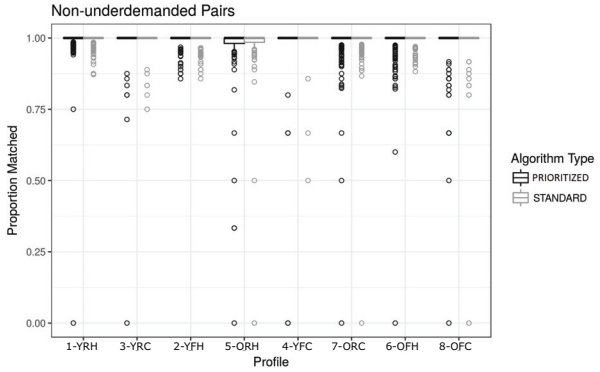

Figure 5: The proportions of overdemanded, self-demanded, and reciprocally demanded pairs grouped together matched over the course of the simulation, by profile type and algorithm type. N = 60 runs were used for each box.

Experiment 3: Transforming Bradley-Terry scores

Experiment One may well wonder whether using the Bradley-Terry scores as weights is well motivated, especially because the difference in scores between the top two profiles is so large. This difference reflects that it is very unlikely that the top profile would not be preferred by a subject, but this does not imply that saving someone of profile 1 is more than four times as important as saving someone of profile 3. Presumably, the ideal weights used in the algorithm

would be monotonically increasing in the BT scores, but it is not clear that they should be proportional. To explore the impact of this on the matchings produced by the prioritizing algorithm, we tried alternative weights, given below.

	Profile							
	1	2	3	4	5	6	7	8
ORIGINAL	1	.103	.236	.036	.070	.011	.024	.003
LINEAR	1	.998	.999	.996	.997	.994	.995	.993

Table 4: Two weight vectors. The first represents the original BT scores as used in PRIORITIZED; the second agrees with the BT scores on the ordering, but the weights are linear in the rank of the profile, as used in LINEAR PRIORITIZED.

The alternative weights result in the profiles being ranked in the same order as the BT scores, but make the difference between sequential weights small and identical. We again ran 20 simulations of 5 simulated years of daily matching, this time comparing the prioritized algorithm using the original BT scores as weights to the prioritized algorithm using the alternative weights. We hypothesized that the profile *ranking* was primarily responsible for the differences in matching and that beyond this, the magnitude of the BT scores would not have a great impact. Hence, since both of these vectors of weights rank profiles the same, we expected them to match profiles in very similar proportions.

Results The proportions of pairs matched using each weight vector are plotted in Figure 6. The matching using the original weights is again called "PRIORITIZED", while the matching using the new weight vector is called "LINEAR PRIORITIZED". The results confirm our hypothesis. There was very little difference in the matchings produced by the PRIORITIZED and LINEAR PRIORITIZED algorithms, and what difference there was could be easily explained by the fact that a slightly different set of pairs enter the pool for each algorithm type. We also tried other weight vectors that assigned different weights to each profile, but that agreed with the initial prioritizing algorithm on the order of the profiles, and found similarly little difference. These results suggest that the profile ranking induced by the weights is primarily responsible for the impact of the prioritizing algorithm, while beyond that varying the weights makes little difference.

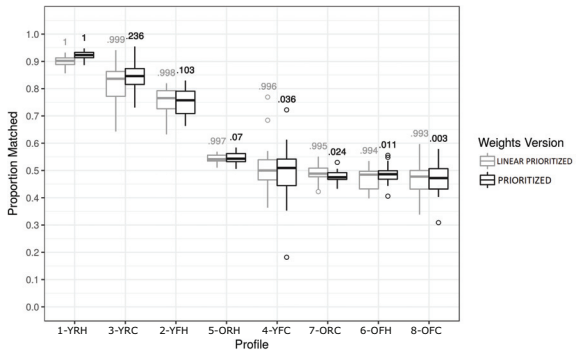

Figure 6: The proportions of underdemanded pairs matched over the course of the simulation, by profile and algorithm. The "PRIORITIZED" algorithm matches using the original profile weights, while the "LINEAR PRIORITIZED" algorithm matches using the alternative weights given above.

Discussion

Our study serves as a proof of concept for the proposed method of soliciting and using prioritization weights, but we do not advocate directly applying the weights obtained in our limited study to a real kidney exchange. For one, a real kidney exchange would require each of the attributes considered to be able to take more possible values than we tested in our mere pairwise comparisons (e.g., there should be more than two values for "age"). Whoever eventually makes the judgments about who should be prioritized (in our study this was left to MTurkers, who may not be representative of the general population) should also have a chance to obtain expert advice—for example, about what the prognosis is for someone with skin cancer in remission. Generally, deploying these techniques in a real kidney exchange should be done with input from representatives of all the stakeholders in such a system—patients, donors, surgeons, other hospital staff, etc. How to best structure the process as a whole is an important topic for future research.

That being said, our work demonstrates that there are no fundamental technical obstacles to building such a system. We have shown one way in which moral judgments can be elicited from human subjects, how those judgments can be statistically modeled, and how the results can be incorporated into the algorithm. We have also shown, through simulations, what the likely effects of deploying such a prioritization system would be, namely that underdemanded pairs would be significantly impacted but little would change for others. We do not make any judgment about whether this conclusion speaks in favor of or against such prioritization, but expect the conclusion to be robust to changes in the prioritization such as those that would result from a more thorough process, as described in the previous paragraph. We also expect the conclusion to hold if the method is applied to real rather than simulated data: while the distribution of donor and patient data in real kidney exchanges is surely different from the simulated one, there are no obvious reasons to suspect that this would change our qualitative conclusion.

Besides being applicable to kidney (and perhaps other organ) exchanges, our study also suggests a roadmap for automated moral decision making in other domains. For example, the idea of obtaining human subjects' judgments to guide AI systems in moral decision making is also being explored for self-driving cars (Bonnefon, Shariff, and Rahwan 2016; Noothigattu et al. 2018). Some aspects of that domain are different. In particular, in that case the need for automated decision-making is driven by the fact that decisions need to be made too fast to be made by a human, whereas in kidney exchanges the need for AI is driven by the fact

that the nature of the search space of all possible matchings makes the problem intractable for a human. Nevertheless, the domains clearly have much in common, and it seems likely that we will be confronted with similar problems in many others. Further research should eventually lead us to a good understanding of best practices for automated moral decision making by generalizing from human judgments.

Acknowledgments This work is partially supported by the project "How to Build Ethics into Robust Artificial Intelligence" funded by the Future of Life Institute, and by NSF IIS-1527434. We thank Lirong Xia, Zhibing Zhao, and Kyle Burris, and members of our moral AI group at Duke, including Yuan Deng, Kenzie Doyle, Jeremy Fox, Max Kramer, and Eitan Sapiro-Gheiler, for feedback on this work.

References

Abraham, D.; Blum, A.; and Sandholm, T. 2007. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *ACM EC*, 295–304.

Anderson, R.; Ashlagi, I.; Gamarnik, D.; and Roth, A. E. 2015. Finding long chains in kidney exchange using the traveling salesman problem. *PNAS* 112(3):663–668.

Ashlagi, I., and Roth, A. E. 2014. Free riding and participation in large scale, multi-hospital kidney exchange. *Theoretical Economics* 9:817–865.

Ashlagi, I.; Gamarnik, D.; Rees, M.; and Roth, A. E. 2017. The need for (long) chains in kidney exchange. Initial version appeared at the ACM Conference on Electronic Commerce (EC-12).

Barnhart, C.; Johnson, E. L.; Nemhauser, G. L.; Savelsbergh, M. W. P.; and Vance, P. H. 1998. Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46(3):316–329.

Biró, P., and Cechlárová, K. 2007. Inapproximability of the kidney exchange problem. *Information Processing Letters* 101(5):199.

Biró, P.; Burnapp, L.; Haase, B.; Hemke, A.; Johnson, R.; van de Klundert, J.; and Manlove, D. 2017. Kidney exchange practices in Europe. First Handbook of the COST Action CA15210: European Network for Collaboration on Kidney Exchange Programmes.

Biró, P.; Manlove, D. F.; and Rizzi, R. 2009. Maximum weight cycle packing in directed graphs, with application to kidney exchange programs. *Discrete Mathematics, Algorithms and Applications* 1(04):499–517.

Blum, A.; Caragiannis, I.; Haghtalab, N.; Procaccia, A.; Procaccia, E.; and Vaish, R. 2017. Opting into optimal matchings. In *SODA*.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Bradley, R. A. 1984. 14 paired comparisons: Some basic procedures and examples. *Handbook of Statistics* 4:299 – 326. Nonparametric Methods.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *AAAI*, 4831–4835. Blue Sky track.

Dickerson, J. P., and Sandholm, T. 2015. FutureMatch: Combining human value judgments and machine learning to match in dynamic environments. In *AAAI*, 622–628.

Dickerson, J. P.; Manlove, D.; Plaut, B.; Sandholm, T.; and Trimble, J. 2016. Position-indexed formulations for kidney exchange. In *ACM EC*.

Dickerson, J. P.; Procaccia, A. D.; and Sandholm, T. 2013. Failure-aware kidney exchange. In *ACM EC*, 323–340.

Elkind, E., and Slinko, A. 2015. Rationalizations of voting rules. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*. Cambridge University Press. Chapter 8.

Ergin, H.; Sönmez, T.; and Ünver, M. U. 2017. Multi-donor organ exchange. Working paper.

Farina, G.; Dickerson, J. P.; and Sandholm, T. 2017. Operation frames and clubs in kidney exchange. In *IJCAI*.

Glorie, K.; van de Klundert, J.; and Wagelmans, A. 2014. Kidney exchange with long chains: An efficient pricing algorithm for clearing barter exchanges with branch-and-price. *Manufacturing & Service Operations Management (MSOM)* 16(4):498–512.

Greene, J.; Rossi, F.; Tasioulas, J.; Venable, K. B.; and Williams, B. C. 2016. Embedding ethical principles in collective decision support systems. In *AAAI*, 4147–4151.

Hajaj, C.; Dickerson, J. P.; Hassidim, A.; Sandholm, T.; and Sarne, D. 2015. Strategy-proof and efficient kidney exchange using a credit mechanism. In *AAAI*, 921–928.

Jia, Z.; Tang, P.; Wang, R.; and Zhang, H. 2017. Efficient near-optimal algorithms for barter exchange. In *AAMAS*, 362–370.

Li, J.; Liu, Y.; Huang, L.; and Tang, P. 2014. Egalitarian pairwise kidney exchange: Fast algorithms via linear programming and parametric flow. In *AAMAS*, 445–452.

Luo, S.; Tang, P.; Wu, C.; and Zeng, J. 2016. Approximation of barter exchanges with cycle length constraints. *CoRR* abs/1605.08863.

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

Mattei, N.; Saffidine, A.; and Walsh, T. 2017. Mechanisms for online organ matching. In *IJCAI*.

Montgomery, R.; Gentry, S.; Marks, W. H.; Warren, D. S.; Hiller, J.; Houp, J.; Zachary, A. A.; Melancon, J. K.; Maley, W. R.; Rabb, H.; Simpkins, C.; and Segev, D. L. 2006. Domino paired kidney donation: a strategy to make best use of live non-directed donation. *The Lancet* 368(9533):419–421.

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; D'Souza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2018. A voting-based system for ethical decision making. In *AAAI*.

O'Neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Rees, M.; Kopke, J.; Pelletier, R.; Segev, D.; Rutter, M.; Fabrega, A.; Rogers, J.; Pankewycz, O.; Hiller, J.; Roth, A.; Sandholm, T.; Ünver, U.; and Montgomery, R. 2009. A nonsimultaneous, extended, altruistic-donor chain. *New England Journal of Medicine* 360(11):1096–1101.

Roth, A. E.; Sonmez, T.; and Unver, M. U. 2004. Kidney exchange. *Quarterly Journal of Economics* 119(2):457–488.

Roth, A.; Sönmez, T.; and Ünver, U. 2005a. A kidney exchange clearinghouse in New England. *American Economic Review* 95(2):376–380.

Roth, A.; Sönmez, T.; and Ünver, U. 2005b. Pairwise kidney exchange. *Journal of Economic Theory* 125(2):151–188.

Tolchinsky, P.; Modgil, S.; Atkinson, K.; McBurney, P.; and Cortés, U. 2012. Deliberation dialogues for reasoning about safety critical actions. *Autonomous Agents and Multi-Agent Systems* 25(2):209.

Toulis, P., and Parkes, D. C. 2015. Design and analysis of multi-hospital kidney exchange mechanisms using random graphs. *Games and Economic Behavior* 91:360–382.

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

Wallach, W., and Allen, C. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.