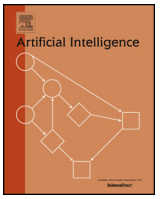
Artificial Intelligence 283 (2020) 103261

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Adapting a kidney exchange algorithm to align with human values ☆

Rachel Freedman [a,*,1], Jana Schaich Borg [b], Walter Sinnott-Armstrong [b], John P. Dickerson [c], Vincent Conitzer [b]

[a] UC Berkeley, United States of America
[b] Duke University, United States of America
[c] University of Maryland, United States of America

ARTICLE INFO

Article history:
Received 27 December 2018
Received in revised form 12 March 2020
Accepted 17 March 2020
Available online 24 March 2020

Keywords:
Moral AI
Human-compatible AI
Computational social choice
Preference aggregation
Kidney exchanges

ABSTRACT

The efficient and fair 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 gets 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.

1. Introduction

As AI is deployed increasingly broadly, AI researchers are confronted with the moral implications of their work. The pursuit of simple objectives, such as minimizing error rates, maximizing resource efficiency, or decreasing response times, often results in systems that have unintended consequences when they confront the real world, such as discriminating against certain groups of people [34]. It would be helpful for AI researchers and practitioners to have a general set of principles with which to approach these problems [45,41,24,16,33].

☆ This paper is an invited revision of a paper which first appeared at the 2018 AAAI Conference on Artificial Intelligence (AAAI-18).

* Corresponding author.
E-mail addresses: rachel.freedman@berkeley.edu (R. Freedman), js524@duke.edu (J.S. Borg), ws66@duke.edu (W. Sinnott-Armstrong), john@cs.umd.edu (J.P. Dickerson), conitzer@cs.duke.edu (V. Conitzer).

[1] Work performed while at Duke University.

https://doi.org/10.1016/j.artint.2020.103261
0004-3702/

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 [13]. 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? [27]), 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 [40]. 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 [19], the United Kingdom [30], the Netherlands [23], and so on [9].

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 would clearly be a morally laden decision. How should this affect the role of the AI researcher developing these systems? From a purely 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. The methodology that we develop integrates this prioritization into our development work.

1.1. 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).

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. For example, we would need to consult domain experts to determine which patient characteristics should be used to determine edge weights. We would also need to involve stakeholders such as policy-makers, doctors, and kidney exchange participants in the process for determining weights. For this reason, we execute this methodology in a limited fashion as a proof-of-concept, and evaluate the results in simulations.

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" [6,42] combination of blood types; for them, their chances rise or drop significantly depending on their tiebreaking weights.

2. 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.

2.1. Graph formulation

In this work, as is standard [40,38,39], 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. For example, a recent report [44] proposes revising the policy for setting edge weights to incorporate the patient's "calculated reactive panel antibody" (which influences their likelihood of finding a match) and whether the pair has previously been a part of a "failed exchange" (in which the donor donates a

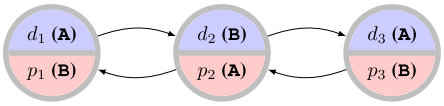


Fig. 1. A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

kidney, but their corresponding patient does not receive one). The policy already includes a multitude of other factors, including which hospital registered the pair and whether the patient has previously donated an organ.[2]

The weight w_e of the edge e from vertex v_i to vertex v_j 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 Fig. 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.

Many fielded kidney exchanges gain great utility through the use of *chains* [32,37,4,5]. 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 [21,22] 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 [6,28,25,12,31].

2.2. 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 [1,11], and is also hard to approximate [10,29,26]. Thus, fielded kidney exchanges use integer program (IP) formulations to solve this difficult combinatorial optimization problem.

The first approach to clearing large kidney exchanges, due to Abraham et al. [1], built a custom branch and price [7] integer program solver; generalizations of, and improvements on, their basic model have addressed scalability issues [23,4, 17,18]. 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.

2.3. Tiebreaking and prioritization: examples

Consider again the compatibility graph given in Fig. 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

[2] For a more detailed look into the inner workings of this process that sets edge weights, we direct the reader to a recent report by the UNOS US-wide kidney exchange [44].

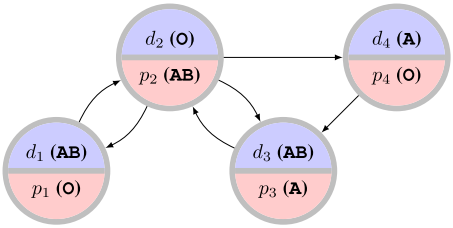


Fig. 2. A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

Table 1
Patient-donor blood-type compatibility. A checkmark denotes compatibility between the patient blood type in the column heading and the donor blood type in the row heading. For example, patients with blood type AB are compatible with all donor blood types, and donors with blood type O are compatible with all patient blood types.

		Patient			
		A	B	AB	O
Donor	A	✓		✓	
	B		✓	✓	
	AB			✓	
	O	✓	✓	✓	✓

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.

Now, consider the graph in Fig. 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. (For a complete description of which patient and donor blood types are compatible, see Table 1.) 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.

3. 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. Because this procedure was intended as a proof-of-concept, we gathered preference data from participants recruited through the online platform Amazon Mechanical Turk ("MTurk").[3] However, if this procedure were used in a real-life kidney exchange, medical experts and other stakeholders would need to be involved in the process of determining weights.[4]

3.1. 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 asked to read a brief description of the kidney transplant waiting list process, and then asked to imagine that a country is developing a new policy for allocating kidneys to patients on the waiting list. Each participant was asked to report four potential patient attributes that they thought the kidney allocation policy "morally ought to take into account," and four attributes that they thought the policy "morally ought NOT to take into account." Each participant received $0.85 as compensation for their participation.

Participants' responses were independently sorted into attribute categories, including those listed in Table 2, by two different researchers. Attributes that the UNOS algorithm already takes into account, such as patient-donor medical compatibility, were discarded. The number of participants who mentioned each of the remaining attributes is noted in Table 2.

[3] All experiments were conducted between fall 2016 and summer 2017.

[4] That being said, it is not immediately clear what the optimal mix of stakeholders would be. For example, it does not seem that medical training is especially helpful for evaluating how important it is whether a patient has dependents, such as small children.

Table 2
Categorized responses to the Attribute Collection Survey. The "Ought" column counts the number of responses in each category that participants thought should be used to prioritize patients. The "Ought NOT" column counts those that participants thought should not be used to prioritize patients. Categories are listed in order of popularity.

Category	Ought	Ought NOT
Age	80	10
Health - Behavioral	53	5
Health - General	44	9
Dependents	18	5
Criminal Record	9	4
Expected Future	8	1
Societal Contribution	7	3
Attitude	6	0

Because we are interested in improving the kidney allocation process, we only included those categories that more survey participants thought *ought* to be taken into account than participants thought *ought not* to be taken into account.

Because participants were asked to propose these attributes themselves, these results reflect which attributes occurred to them during the survey. This may skew the results in favor of attributes that seem more directly relevant to the medical context. For example, it's possible that 30 of the survey participants would have answered "yes" if directly asked whether criminal record should be taken into account, but because this aspect of personal life is not clearly related to health, only a few of those thought of it when prompted to consider public health policy during the survey. Additionally, we explicitly listed age as an example of the sort of attribute the policy might consider, which likely biased participants toward including it in their responses. We chose to prime with age in order to direct responses toward the sort of specific, individual attributes that the revised policy might take into account. We chose age specifically because it is a common response to informal iterations of this survey, and in fact is already included in current kidney allocation policy to a certain extent, so we hoped that the skewing effects of the priming would be minimal.

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. The least-commonly reported categories were "Criminal Record", "Expected Future", which included responses about patients' future life expectancy and expected quality of life post-surgery, "Societal Contribution", and "Attitude", which included responses about patients' psychological state and mental preparation for the surgery and recovery process.

3.1.1. Participant demographics

The survey participants were very diverse, ranging in age from 22 to 64 (with an average age of 40), ranging in self-reported political views from "extremely liberal" to "extremely conservative", and ranging in educational achievement from "some high school, no diploma" to "doctorate degree". Participants took between 2 and 26 minutes to complete this survey, with an average completion time of 9 minutes.

3.2. 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 3 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 can contribute 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 and is unrelated to kidney disease.

Because there are three binary attributes, there are eight possible patient profiles. These eight unique patient profiles were enumerated and assigned ID numbers. For expositional ease, in this paper, we refer to profiles in text as a combination of $\{Y, O\}$, $\{R, F\}$, and $\{C, H\}$, representing {Young, Old}, {Rare, Frequent} alcohol consumption, and {Cancer, Healthy} status. For example, profile YRH reads:

Table 3
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".

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 4
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.

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%

Patient W.A. is 30 years old, had 1 alcoholic drink per month (prior to diagnosis), and has no other major health problems.

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.

3.2.1. Summary of responses

Aggregate responses to the Kidney Allocation Survey are summarized in Table 4. The "Preferred" column reports the percentage of times that each profile was chosen in all the comparisons in which it appeared.

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. This aligns with responses to our first survey, in which more participants gave responses in the "Health - Behavioral" category than gave responses in the "Health - General" category (see Table 2). (Of course, this observation may not generalize to other health-behavioral and health-general attributes, such as drinking soda and skin cancer that's not in remission.)

3.2.2. Participant demographics

Again, the survey participants were very diverse. They ranged in age from 19 to 70 (with an average age of 37), and again ranged in self-reported political views from "extremely liberal" to "extremely conservative", and in educational achievement from "some high school, no diploma" to "doctorate degree". Participants took between 1.5 minutes and 38.5 minutes to complete this survey, with an average completion time of 7 minutes.

3.3. 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 [14]. 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:

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

To illustrate this model, imagine that individuals a, b, and c are patients waiting for kidney transplants. For each pair of patients, imagine that we have asked 100 survey participants to pick one to receive a kidney. Assume that patient a was picked over patient b 63 times and picked over patient c 72 times, and that patient b was picked over patient c 58 times. There were no ties. We can use the Bradley-Terry model to estimate a score representing the value that survey participants place on giving a kidney to each patient.

Judging from this sample of 300 comparisons, the probability of patient a being chosen over patient b is $63/100 = 0.63$, the probability of patient a being chosen over patient c is $72/100 = 0.72$, and the probability of patient b being chosen over patient c is $58/100 = 0.58$. Therefore, we have:

$$P(a > b) = 0.63 = \frac{p_a}{p_a + p_b}$$
$$P(a > c) = 0.72 = \frac{p_a}{p_a + p_c}$$
$$P(b > c) = 0.58 = \frac{p_b}{p_b + p_c}$$

When we fit the model to these results and assign score 1.00 to p_a, p_b is estimated as 0.57, and p_c is estimated as 0.40. It is important to note that these scores are only meaningful relative to each other. In particular, scaling all the scores p_i by the same factor would not affect the predictions. Based on these scores, the preference probabilities are estimated as follows:

$$P(a > b) = \frac{p_a}{p_a + p_b} = \frac{1.00}{1.00 + 0.57} \approx 0.64$$
$$P(a > c) = \frac{p_a}{p_a + p_c} = \frac{1.00}{1.00 + 0.40} \approx 0.71$$
$$P(b > c) = \frac{p_b}{p_b + p_c} = \frac{0.57}{0.57 + 0.40} \approx 0.59$$

Note that these scores do not exactly line up with the empirical fractions with which each patient is chosen (0.63, 0.72, and 0.58, respectively); this is because we only have 2 degrees of freedom. Specifically, if we decrease p_a then the estimate gets closer to the first empirical fraction but further away from the second; if we decrease p_b the estimate gets closer to the third but further away from the first; and if we decrease p_c the estimate gets closer to the second but further away from the third.

The BT scores (that we estimate based on our data) 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 [24,16,33], 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 [20]).

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 Table 5 below.

3.4. Adapting the algorithm

The final step was to incorporate the obtained weights into the kidney exchange market clearing algorithm. Because our human subject 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.

Table 5
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.

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

To find a matching, our adapted (prioritized) algorithm first runs the basic IP-based algorithm due to Abraham et al. [1] with unit edge weights (i.e., $w_e = 1 \ \forall e \in E$). Given a pool of patient-donor pairs, this algorithm returns a set of kidney exchange cycles that maximizes the number of patients who receive a kidney without regard to their personal characteristics (other than medical compatibility). Our algorithm records the number of patients who 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, $type : V \to \{1, \ldots, 8\}$ mapping a vertex to its patient's profile, and w_θ denoting the score of profile θ, we solve:

$$\begin{array}{lll} \max & \sum_{c \in C(L,K)} \left[\sum_{(u,v) \in c} w_{type(v)} \right] x_c & \\ \text{s.t.} & \sum_{c: v \in c} x_c \leq 1 & \forall v \in V \\ & \sum_{c \in C(L,K)} |c| x_c \geq Q & \\ & x_c \in \{0, 1\} & \forall c \in C(L, K) \end{array}$$

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.

4. 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.

4.1. Experimental setup

Based on previously developed tools [19], we built a simulator to mimic daily matching in a real-world kidney exchange pool.[5] 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. Not all of the matched pairs are executed, because in real-life situations many algorithmic matches fail to go to transplant due to last-minute medical incompatibilities, surgeons rejecting a donor organ, or other logistical difficulties [4,18] We model this by executing matches with a probability of 0.5 instead of 1. The demographics of our simulated pool were designed to reflect the UNOS kidney exchange pool where possible, and otherwise the general US population.

4.2. Experiment 1: matchings with pair scores

4.2.1. 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.

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

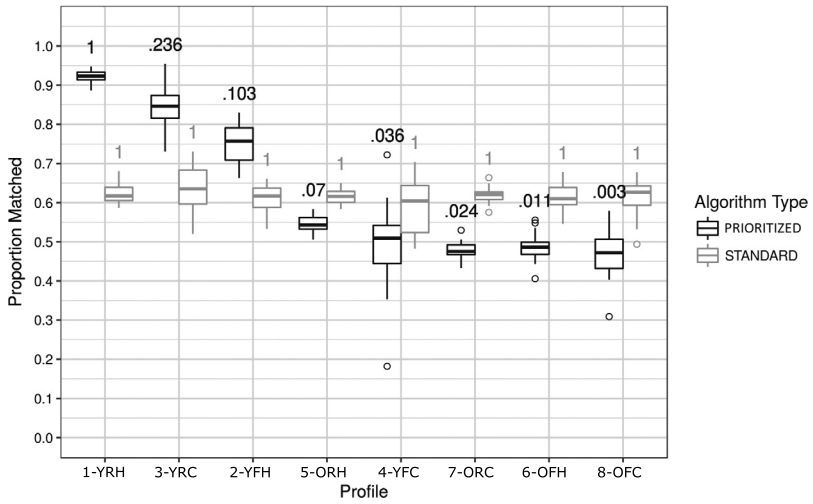


Fig. 3. The proportions of pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally, 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 × the interquartile range of the median, and the small circles denote outliers beyond this range.

4.2.2. Hypothesis

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.

4.2.3. Results

The proportions of pairs of each profile type matched by the original and prioritizing algorithms are plotted in Fig. 3 above. "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 Fig. 3, matched pairs approximately 62% of the time, regardless of their profile, while the prioritizing algorithm, called "PRIORITIZED" in Fig. 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 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.

4.3. Experiment 2: matchings evaluated by blood type

4.3.1. Experiment

Blood type is a major factor in determining patient-donor biological compatibility (see Table 1 for a summary of blood type compatibility). Patients with difficult-to-match blood types are more likely to struggle to find a compatible donor, and consequently can be disproportionately represented in kidney exchange pools. To explore how the modified algorithm treats patients with these blood types, 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 [6,42]. *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.

4.3.2. Hypothesis

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

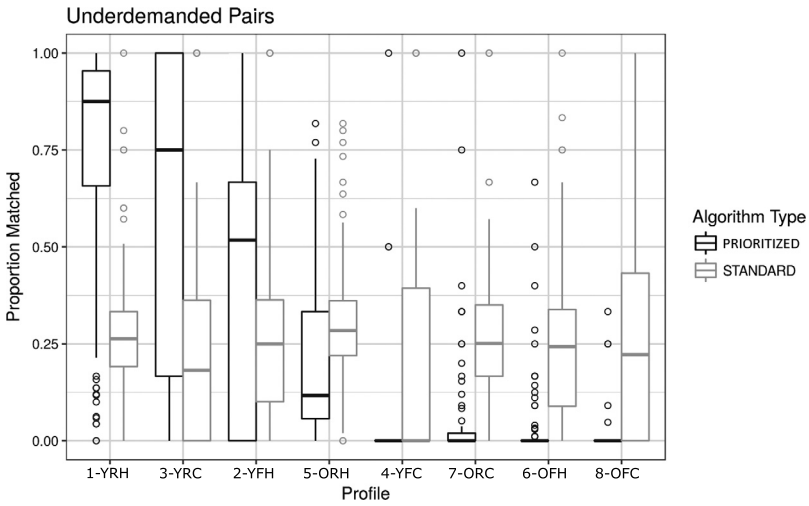


Fig. 4. The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

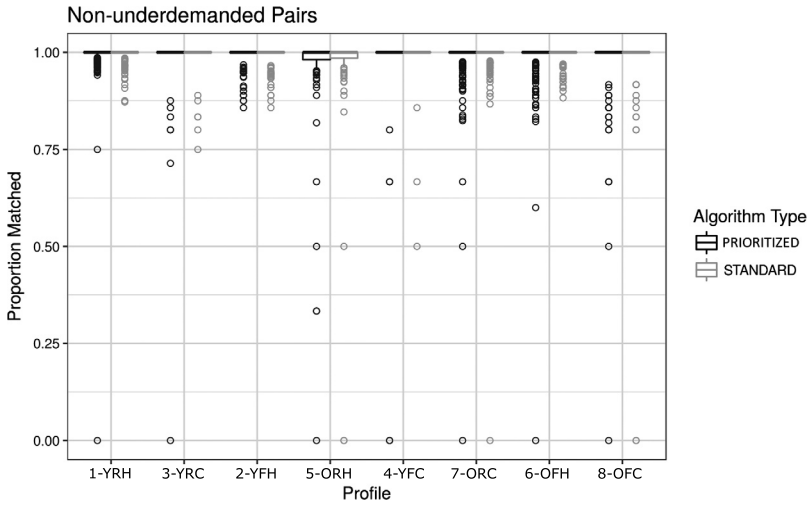


Fig. 5. The proportions of overdemanded, self-demanded, and reciprocally demanded pairs grouped together matched over the course of the simulation, by profile type and algorithm type. N = 60 runs were used for each box.

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.

4.3.3. Results

The results confirm our hypothesis. The proportions of underdemanded pairs matched are plotted in Fig. 4. We found the proportions of overdemanded, self-demanded, and reciprocally demanded profiles matched to be fairly similar, so we grouped them together in Fig. 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.

4.4. Experiment 3: transforming Bradley-Terry scores

4.4.1. 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

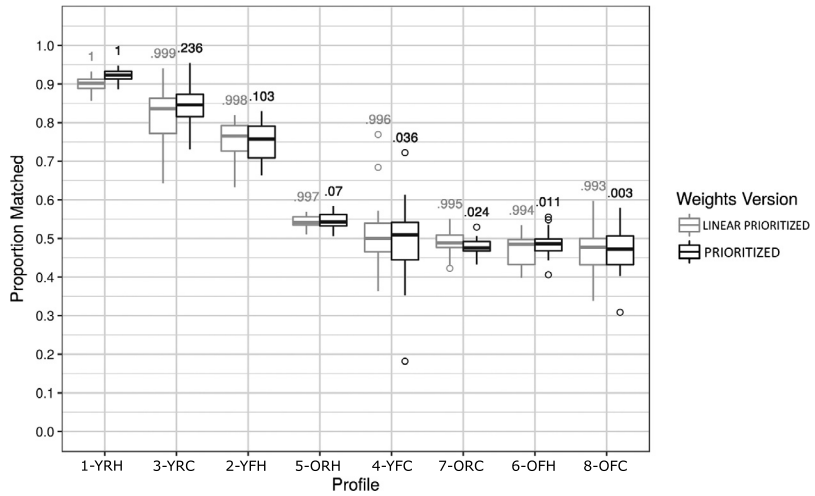


Fig. 6. The proportions of underdemanded pairs matched over the course of the simulation, by profile and algorithm. The "PRIORITIZED" algorithm matches using the original profile weights, while the "LINEAR PRIORITIZED" algorithm matches using the alternative weights given above.

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 the sizes of the gaps between the weights on the matchings produced by the PRIORITIZED algorithm, we tried alternative weights, given below (Table 6).

Table 6
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.

	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

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.

4.4.2. Hypothesis

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.

4.4.3. Results

The proportions of pairs matched using each weight vector are plotted in Fig. 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.

5. Discussion

In this section, we discuss the potential for applying these results to real-world kidney exchanges, some of the ethical context of our work, and directions for future research.

5.1. Application in real kidney exchanges

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.

5.2. Artificial morality

Our work is also a concrete proof-of-concept of a hybrid approach to artificial morality. This hybrid combines and contrasts with both top-down approaches and bottom-up approaches [2,46].

Top-down approaches provide a computer with a general ethical theory along with facts that are morally relevant according to the theory. The machine then infers moral judgments or makes moral decisions by applying the theory to the facts. This top-down approach must begin by choosing a moral theory to program into computers. The problem is that ethicists support a wide variety of moral theories, and it is hard to see how to justify insisting on one theory instead of another. The second problem for the top-down approach is that such theories are too vague to implement, they conflict in real-life decisions, and they can yield disasters [2,46,3,35,15]. It might seem innocuous for a computer to follow a rule like "Minimize harm," but what if a computer decides that killing all humans will minimize harm in the long run?

Bottom-up approaches try to avoid assuming any moral theory by using machine learning trained on human descriptions of concrete moral problems to predict human moral judgments. This system mirrors one way in which children learn morality, so it resembles Alan Turing's original proposal for developing artificial intelligence. Over 50 years ago, when faced with the problem of developing an artificial agent capable of making decisions like an adult human, Turing presciently suggested, "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education, one would obtain the adult brain." [43] It is not completely clear how children learn morality, but one element involves encounters with concrete moral problems. Accordingly, a purely bottom-up approach might try to build AI systems that can learn solely by exposing them to moral examples.

In order to learn in this way, an artificial moral judge or agent must be able to safely extrapolate what it learned in some cases to novel environments that it may face in the future. This ambitious purpose requires a data set that is large and varied enough to teach genuine underlying moral principles that are projectible into these new environments as opposed to narrowly applicable surface features that predict human moral judgments only in the training set. Moreover, we want to know not only that an act is wrong but also why it is wrong—what makes it wrong. Otherwise, we cannot give comprehensible justifications for controversial decisions, and we have no way to check on whether or when the system is working properly. These goals are difficult or impossible to achieve within uninterpretable ML systems [8]. Moreover, today's AI systems lack a broad/commonsense understanding of our (human) world, and it seems that such an understanding would be a necessary component of a system that could make moral judgments across a broad range of settings. This task could be aided by computational models of case-based and value-based reasoning and argumentation. For a variety of approaches, see Rahwan and Simari [36].

While top-down and bottom-up approaches both contain promising elements for developing moral artificial intelligence, each approach also faces serious challenges [8]. Their flaws suggest that we might be able to do better by combining the two approaches into unified systems that achieve the benefits without the problems of each [16]. Top-down supervision and organization can enable artificial agents to justify their decisions in terms of moral principles that are comprehensible to humans, while bottom-up learning has the potential to deal with complex facts in particular cases.

Our particular hybrid attempts to reduce the arbitrariness of top-down approaches by crowd-sourcing a list of features that humans see as morally relevant. Humans do not include some characteristics, such as shirt color, as relevant to kidney

exchanges, and they are responsible for determining which patient characteristics are important. These features of alternatives can then be used both to constrain the data and also to provide an interpretable basis for the algorithm's predictions. In this way, our hybrid introduces some minimal theory in the form of morally relevant features in order to solve the main problems of competing top-down and bottom-up approaches.

This hybrid method is particularly well-suited to developing ethical machine reasoning in constrained domains where it is clear which features of acts are morally relevant. In such a domain, it is possible to create models of multiple individuals' moral decisions, and then to have these models *vote* over what the right decision is overall [16]. Noothigattu et al. [33] recently applied a version of this approach to ethical decision-making for autonomous vehicles. They aggregated human moral judgments about autonomous vehicles colliding with, for example, a pedestrian (likely killing them) or a wall (likely killing the driver). Their method assumes that causing death along with a few other features are morally relevant.

We applied a similar approach to kidney allocation in this paper. Our hybrid approach is not without its own challenges, however. One is that human moral judgments are inconsistent within and across individuals, so a machine learning system can at best predict a subset (though perhaps a majority) of human moral judgments. We will need to decide how to make social decisions in light of such disagreements.

Moreover, humans often exhibit biases, such as racial and gender discrimination, that they themselves reject as improper and would want an artificial moral agent to avoid. This problem can be reduced (though not fully solved) by designing the artificial intelligence system to include only features that most humans deem to be morally relevant. If we had included the characteristic "race" in our patient descriptions, the algorithm might have learned to take race into account. Leaving out that characteristic avoids this undesirable result, though it still leaves open the possibility of more subtle and hidden forms of bias. These problems for our hybrid approach will be the topic of future work.

5.3. Future research

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 [13,33]. 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is partially supported by the project "How to Build Ethics into Robust Artificial Intelligence" funded by the Future of Life Institute (grants 2016-158697 and 2017-174867), by the Templeton World Charity Foundation grant TWCF0321, by NSF IIS-1527434, and by Duke Bass Connections. Dickerson was supported in part by NSF CAREER Award IIS-1846237 and a Google Faculty Research Award. Conitzer was supported in part by NSF IIS-1814056. 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

[1] D. Abraham, A. Blum, T. Sandholm, Clearing algorithms for barter exchange markets: enabling nationwide kidney exchanges, in: ACM EC, 2007, pp. 295–304.
[2] C. Allen, I. Smit, W. Wallach, Artificial morality: top-down, bottom-up, and hybrid approaches, Ethics Inf. Technol. 7 (3) (2005) 149–155.
[3] M. Anderson, S. Anderson, Machine Ethics, Cambridge Univ. Press, 2011, pp. 231–492.
[4] R. Anderson, I. Ashlagi, D. Gamarnik, A.E. Roth, Finding long chains in kidney exchange using the traveling salesman problem, Proc. Natl. Acad. Sci. 112 (3) (2015) 663–668.
[5] I. Ashlagi, D. Gamarnik, M. Rees, A.E. Roth, The need for (long) chains in kidney exchange, 2017, initial version appeared at the ACM Conference on Electronic Commerce (EC-12).
[6] I. Ashlagi, A.E. Roth, Free riding and participation in large scale, multi-hospital kidney exchange, Theor. Econ. 9 (2014) 817–865.
[7] C. Barnhart, E.L. Johnson, G.L. Nemhauser, M.W.P. Savelsbergh, P.H. Vance, Branch-and-price: column generation for solving huge integer programs, Oper. Res. 46 (3) (1998) 316–329.
[8] K. Baum, H. Hermanns, T. Speith, From machine ethics to machine explainability and back, in: ISAIM, 2018.
[9] P. Biró, L. Burnapp, B. Haase, A. Hemke, R. Johnson, J. van de Klundert, D. Manlove, Kidney exchange practices in Europe, in: First Handbook of the COST Action CA15210: European Network for Collaboration on Kidney Exchange Programmes, 2017.
[10] P. Biró, K. Cechlárová, Inapproximability of the kidney exchange problem, Inf. Process. Lett. 101 (5) (2007) 199.

[11] P. Biró, D.F. Manlove, R. Rizzi, Maximum weight cycle packing in directed graphs, with application to kidney exchange programs, Discrete Math. Algorithms Appl. 1 (04) (2009) 499–517.
[12] A. Blum, I. Caragiannis, N. Haghtalab, A. Procaccia, E. Procaccia, R. Vaish, Opting into optimal matchings, in: SODA, 2017.
[13] J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles, Science 352 (6293) (2016) 1573–1576.
[14] R.A. Bradley, 14 paired comparisons: some basic procedures and examples, in: Nonparametric Methods, in: Handbook of Statistics, vol. 4, 1984, pp. 299–326.
[15] S. Bringsjord, G. Naveen Sundar, B.F. Malle, M. Scheutz, Contextual deontic cognitive event calculi for ethically correct robots, in: International Symposium on Artificial Intelligence and Mathematics, ISAIM 2018, Fort Lauderdale, Florida, USA, January 3-5, 2018, 2018.
[16] V. Conitzer, W. Sinnott-Armstrong, J.S. Borg, Y. Deng, M. Kramer, Moral decision making frameworks for artificial intelligence, in: AAAI, 2017, pp. 4831–4835, Blue Sky track.
[17] J.P. Dickerson, D. Manlove, B. Plaut, T. Sandholm, J. Trimble, Position-indexed formulations for kidney exchange, in: ACM EC, 2016.
[18] J.P. Dickerson, A.D. Procaccia, T. Sandholm, Failure-aware kidney exchange, Manag. Sci. 65 (4) (2019) 1768–1791.
[19] J.P. Dickerson, T. Sandholm, FutureMatch: combining human value judgments and machine learning to match in dynamic environments, in: AAAI, 2015, pp. 622–628.
[20] E. Elkind, A. Slinko, Rationalizations of voting rules, in: F. Brandt, V. Conitzer, U. Endriss, J. Lang, A.D. Procaccia (Eds.), Handbook of Computational Social Choice, Cambridge University Press, 2015 (Chapter 8).
[21] H. Ergin, T. Sönmez, M.U. Ünver, Multi-donor organ exchange, Working Paper, 2017.
[22] G. Farina, J.P. Dickerson, T. Sandholm, Operation frames and clubs in kidney exchange, in: IJCAI, 2017.
[23] K. Glorie, J. van de Klundert, A. Wagelmans, Kidney exchange with long chains: an efficient pricing algorithm for clearing barter exchanges with branch-and-price, Manuf. Serv. Oper. Manag. 16 (4) (2014) 498–512.
[24] J. Greene, F. Rossi, J. Tasioulas, K.B. Venable, B.C. Williams, Embedding ethical principles in collective decision support systems, in: AAAI, 2016, pp. 4147–4151.
[25] C. Hajaj, J.P. Dickerson, A. Hassidim, T. Sandholm, D. Sarne, Strategy-proof and efficient kidney exchange using a credit mechanism, in: AAAI, 2015, pp. 921–928.
[26] Z. Jia, P. Tang, R. Wang, H. Zhang, Efficient near-optimal algorithms for barter exchange, in: AAMAS, 2017, pp. 362–370.
[27] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions, Working Paper 23180, National Bureau of Economic Research, 2017.
[28] J. Li, Y. Liu, L. Huang, P. Tang, Egalitarian pairwise kidney exchange: fast algorithms via linear programming and parametric flow, in: AAMAS, 2014, pp. 445–452.
[29] S. Luo, P. Tang, C. Wu, J. Zeng, Approximation of barter exchanges with cycle length constraints, CoRR, arXiv:1605.08863, 2016.
[30] D. Manlove, G. O'Malley, Paired and altruistic kidney donation in the UK: algorithms and experimentation, ACM J. Exp. Algorithmics 19 (1) (2015).
[31] N. Mattei, A. Saffidine, T. Walsh, Mechanisms for online organ matching, in: IJCAI, 2017.
[32] R. Montgomery, S. Gentry, W.H. Marks, D.S. Warren, J. Hiller, J. Houp, A.A. Zachary, J.K. Melancon, W.R. Maley, H. Rabb, C. Simpkins, D.L. Segev, Domino paired kidney donation: a strategy to make best use of live non-directed donation, Lancet 368 (9533) (2006) 419–421.
[33] R. Noothigattu, S.N.S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, A.D. Procaccia, A voting-based system for ethical decision making, in: AAAI, AAAI Press, 2018, pp. 1587–1594.
[34] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Broadway Books, 2017.
[35] L.M. Pereira, A. Saptawijaya, Agent morality via counterfactuals in logic programming, in: Proceedings of the Bridging@CogSci Workshop, London, UK, 2017, pp. 39–53.
[36] I. Rahwan, G.R. Simari, Argumentation in Artificial Intelligence, vol. 47, Springer, 2009.
[37] M. Rees, J. Kopke, R. Pelletier, D. Segev, M. Rutter, A. Fabrega, J. Rogers, O. Pankewycz, J. Hiller, A. Roth, T. Sandholm, U. Ünver, R. Montgomery, A nonsimultaneous, extended, altruistic-donor chain, N. Engl. J. Med. 360 (11) (2009) 1096–1101.
[38] A. Roth, T. Sönmez, U. Ünver, A kidney exchange clearinghouse in New England, Am. Econ. Rev. 95 (2) (2005) 376–380.
[39] A. Roth, T. Sönmez, U. Ünver, Pairwise kidney exchange, J. Econ. Theory 125 (2) (2005) 151–188.
[40] A.E. Roth, T. Sonmez, M.U. Unver, Kidney exchange, Q. J. Econ. 119 (2) (2004) 457–488.
[41] P. Tolchinsky, S. Modgil, K. Atkinson, P. McBurney, U. Cortés, Deliberation dialogues for reasoning about safety critical actions, Auton. Agents Multi-Agent Syst. 25 (2) (2012) 209.
[42] P. Toulis, D.C. Parkes, Design and analysis of multi-hospital kidney exchange mechanisms using random graphs, Games Econ. Behav. 91 (2015) 360–382.
[43] A.M. Turing, Computing machinery and intelligence, Mind 59 (236) (1950) 433–460.
[44] UNOS, Revising kidney paired donation pilot program priority points, OPTN/UNOS Public Comment Proposal, 2015.
[45] W. Wallach, C. Allen, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, 2008.
[46] W. Wallach, C. Allen, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, Inc., New York, NY, USA, 2010, pp. 83–98 (Chapter 6).