COV886 Project
Using Computational Social Choice for Moral Decision Making

Sarthak Singla, Devesh Pant

April 10, 2022

Autonomous decision-making systems frequently confront a choice between two or more alternatives. The criteria for comparing the alternatives are often unclear or unethical. In this report, we discuss two such problems [1]. The first is the Kidney Exchange Problem, where a central market maker allocates living kidney donors to patients in need of an organ. The allotment chosen from the numerous may determine a person's fate. The second is the Trolley Problem in Autonomous Vehicles. A brake failure may encompass deciding whom to kill and whom to save. This report describes a general principle for ethical decision making in autonomous systems using voting.

1 Introduction

Artificial intelligence-based systems are increasingly getting integrated into our daily lives. Numerous decisions made by an autonomous system can significantly impact human lives, either directly or indirectly. Consider a hypothetical scenario in which many people attempt to flee a location simultaneously due to a terrorist attack. A centrally deployed transportation-cost algorithm may detect an increase in transportation out of the area and dramatically increase its cost to reduce demand, increase supply, and increase profit. Simple objective functions, therefore, may result in undesirable outcomes. Such effects are ethically reprehensible.

[1] presented in (Freedman et al., 2020) and (Noothigattu et al., 2017). Our discussion will focus on (Freedman et al., 2020), but we will compare and contrast it with (Noothigattu et al., 2017).

An alternative may be to decouple moral decision making from AI systems. However, moral decision making cannot be left to humans in a computational system which

- Cannot be easily separated from the moral decision making part and solving, which exceeds human capabilities. For instance, the Kidney Exchange Problem discussed in (Freedman et al., 2020) and Section 3.
- Needs to make split-second moral decisions. For instance, the trolley problem is discussed in (Noothigattu et al., 2017) and Section 4.
- Is massively deployed, with millions of moral decisions that add up to something significant. For instance, the computational systems behind online advertisements.

Future machines will increasingly rely on AI, with examples being organ matching systems, self-driving cars, collateral damage estimators[2] (IHL In Action, 2021). In these scenarios, the machines must adhere to ethical principles aligned with human moral values.

Formal ground truth ethical principles have been debated upon for centuries by ethicists, and their absence poses a major challenge towards automating ethical decisions. Humans differ in their moral judgments, and there is no obvious answer as to who is correct. In the lack of ground-truth ethical standards, Dwork et al. suggested that we must rely on an "approximation as agreed upon by society" (Dwork et al., 2011). However, the question then becomes to find the approximation.

Recent papers by (Greene et al., 2016) and (Conitzer et al., 2017) argue that the field of computational social choice, which works with algorithms for aggregating individual preferences towards collective outcomes, may give tools for ethical decision making. In particular, (Conitzer et al., 2017) raise the possibility of "letting our models of multiple people's moral values vote over the relevant alternatives.", thereby giving a window to automate ethical decision making.

This report discusses two crucial problems that use **Computational Social Choice** to develop a model for aggregating society's views on moral dilemmas. The first one is the **clearing house problem in Kidney Exchange**. In a kidney exchange, patients who need a kidney transplant and have a willing but the incompatible live donor may attempt to trade their donors' kidneys. The moral decision aspect is the prioritization of some patients over the others to select one solution in the presence of multiple possible.

The second problem is the **trolley problem in autonomous vehicles**: An autonomous vehicle has a brake failure, leading to unavoidable catastrophic outcomes, however, the vehicle's AI can make an informed decision. Should it stay its course and hit a wall, killing its passengers, one of whom is a young girl? Or swerve and kill a male athlete and his dog, crossing the street on a red light? We will mainly discuss the first problem and then extend its framework to the second problem.

[2] An AI model used during wars to determine best targets, weighing civilian casualties, damage to essential infrastructure, military targets and military infrastructure

2 Related Work

The problem of ethical decision making in AI has recently gained widespread popularity. It has been the focus of several recent research works (Wallach and Allen, 2009; Greene et al., 2016; Conitzer et al., 2017)

In particular, the first study that integrates Computational Social Choice with Moral Decision making is that of (Greene et al., 2016). Their paper discusses the potential of modelling ethical principles as a "dummy" agent's preferences as part of a more extensive system. A more recent paper on ethical decision making in AI by (Conitzer et al., 2017) addresses a variety of frameworks, including game-theoretic models, social choice, and machine learning. It states that by aggregating the views of multiple humans, mistakes made by any individual fade in the aggregate and result in a better system ethically.

3 Paper 1: Adapting a Kidney Exchange algorithm to align with human values

Worldwide, many people suffer from chronic renal disease. Every year, about 2 million people die of renal failure (ScienceDaily, 2015). India alone accounts for more than 19% of all global deaths per year (The Lancet Global Health, 2017). Patients currently have two treatment options: dialysis or kidney transplantation.

Dialysis is beneficial for a brief period, but it does have some disadvantages. It is often conducted twice a week for four to five hours, which is inconvenient and exhausting for the patient. Additional concerns include dialysis's cost and low mortality rate.

The alternative is a kidney transplant, a surgical procedure that requires a compatible donor. Only when the blood types of the donor and recipient are compatible can a kidney transplant be performed (Table 3.1). A reasonably large market exists for kidney exchange, in which a central body matches living kidney donors with patients in need. There is a severe donor shortage due to many patients requiring kidney transplantation. The waiting list in the kidney exchange market in the United States is approximately 90,000, whereas it is over 200,000 in India (Organ Procurement and Transplantation, 2022; The Narayana Health, 2019).

A kidney exchange uses an algorithm to pick who gets a kidney. Patients and donors may be prioritised depending on weights decided ad hoc by a central authority. The central authority has the power to choose one over the other, and it may do so because of some incentive and not a reasonable justification. From the patient's standpoint, the allocation of kidneys has life-death implications. This paper presents a comprehensive methodology consistent with human moral values for determining the weights associated with individual patient profiles in kidney exchange.

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

Table 3.1: Blood-type compatibility between donors and patients

3.1 Kidney Exchange Model

A Kidney Exchange Problem can be modelled as a directed compatibility graph G(V,E) as shown in Figure 3.1. A **vertex** in the graph represents a donor-patient pair. The **edge** from a vertex u to v represents that the patient of vertex v is compatible and willing to receive the kidney from the donor of vertex u. The **weight** w_i of the edge represents the utility of the vertex v_j obtaining v_i's kidney.

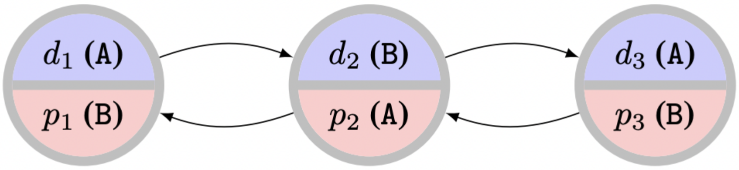

Figure 3.1: A compatibility graph with three patient-donor pairs and two possible cylces.

In general a solution to the kidney exchange problem may consist of Cycles and Chains.

Definition 1 (Cycle)**.** *A cycle in the graph G is sequence of transplants where each vertex receives a kidney from it's previous vertex.*

Definition 2 (Chain)**.** *A chain is a sequence of transplants starting with an altruist donor unconditionally donating their kidney to a patient whose donor donates to another patient, continuing the sequence of transplants.*

For uniformity, altruistic donors are represented as a normal patient-donor pair with a 'dummy' patient.

3.2 The Clearing House Problem

Definition 3 (Matching)**.** *Matching is a set of disjoint cycles and chains in the compatibility graph.*

Cycles and chains need to be disjoint because a living donor can donate only one kidney. Due to logistical issues, solutions with small maximum cycle and chain lengths are preferred.

Definition 4 (Legal Matching)**.** *A legal matching is one which obeys the restrictions on maximum sizes of chains and cycles.*

Definition 5 (Utility)**.** *Utility u is a function from the set of all matchings($\mathcal{M}$) to real numbers($\mathbb{R}$), i.e.*

$$u : \mathcal{M} \to \mathbb{R}$$

Definition 6 (Clearing House Problem)**.** *For the set of all legal matchings $\mathcal{M}$, the clearing house problem is to find a matching M^* that maximizes the utility. Formally:*

$$M^* \in \underset{M \in \mathcal{M}}{argmax}\ u(M)$$

The common utility function used is called the *utilitarian* utility function, and is defined as:

$$u(M) = \sum_{c \in M} \sum_{e \in c} w_e$$

The clearing house problem with limit on maximum cycle length $L > 2$ is NP-hard (Abraham et al., 2007; BIRÓ et al., 2009) and also hard to approximate (Biró and Cechlárová, 2007) Therefore, kidney exchange solvers use Integer Program(IP) based formulations.

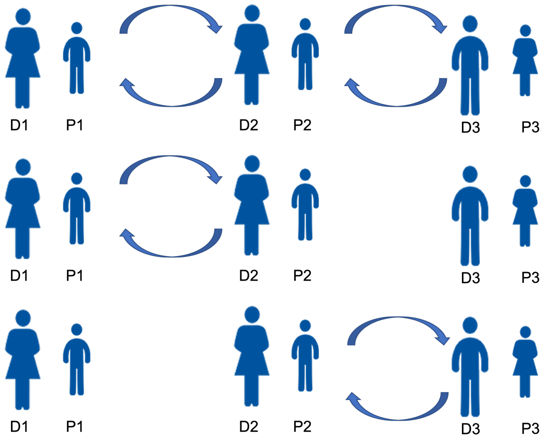

Figure 3.2: An example of the kidney exchange problem

3.3 IP model of Abraham et al.

The paper uses a model similar to the model built by (Abraham et al., 2007) which uses custom branch and price integer solver.

Let $C(L,K)$ denote the set of all cycles with length atmost L and chains with length atmost K. For every $c \in C(L,K)$, let $x_c = \{0,1\}$ and $w_c = \sum_{e \in c} w_e$ then for the Kidney Exchange Problem we have the integer program:

$$max \sum_{c \in C(L,K)} w_c x_c, \text{ such that } \sum_{c : v \in c} x_c \le 1 \quad \forall v \in V.$$

Here, the second constraint accounts for the fact that the cycles and chains must be disjoint.

3.4 The Problem

Consider the following example of the Kidney Exchange Problem shown in Figure 3.2. It can have two solutions. Only one of the leftmost and the rightmost patient-donor pairs can receive a kidney. It rests on the Kidney Exchange committee to decide which solution is chosen. The Abraham et al. algorithm utilizes weights to prioritize one solution over the other, but there is no specification on how to arrive at the weights.

The current paper provides an end-to-end framework to calculate these weights. It compares the results with the baseline being the IP algorithm of Abraham et al. run with unit edge weights.

3.5 Procedure

The authors first utilise a survey to determine the attributes to classify the profiles. The authors then use another survey to find the pairwise comparisons among different patient profiles and find a utility corresponding to each profile. The paper first runs the basic algorithm with all edge weights equal to 1 and finds a solution with maximum cardinality. Then, it feeds the utilities calculated as weights in the IP algorithm and modifies its objective to produce a maximum cardinality solution (calculated in the previous run). The solution obtained is accepted.

3.5.1 Survey 1: Selecting Attributes

A survey was undertaken on **Amazon Mechanical Turk** to get a list of patient attributes deemed ethical by the society to prioritize patients. An open-ended survey, requiring the participants to provide the attributes themselves, was conducted to remove experimenter bias. One hundred participants were asked to read a brief overview of the kidney exchange waiting list procedure and imagine that their country is implementing a new kidney allocation policy. Each participant was asked to provide four characteristics that they thought the kidney distribution policy “should take into account morally” and four that they thought the policy “should not take into account morally.”

The researchers classified survey responses into the categories displayed in Table 3.2, specifically:

- “Age”
- “Health – Behavioral” (controllable aspects of health like diet and drug use)
- “Health – General” (involuntary health issues and other diseases unrelated to kidney disease, such as cancer prognosis)

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

Table 3.2: The Attribute Collection Survey Results. The "Ought" column counts the number of responses in each category from participants who believe it should be used to prioritise patients. Participants who thought they should not be used to prioritise patients were counted in the "Ought NOT" column.

- "Dependents"
- "Criminal Record"
- "Expected Future"(included responses related to the patient's life expectancy post-surgery)
- "Societal Contribution"
- "Attitude"(tells about the patient's psychological state and preparation for surgery and recovery)

Responses that included features already considered, such as medical compatibility, were disregarded.

The top three attributes were: "Age", "Health – Behavioral", and "Health – General". Only these were picked for the next step of the process due to the substantial fall in support between the third most popular and fourth most popular categories.

The authors look into the likelihood of the open survey skewing the acquired qualities towards medical features. For example, given the setting, most participants may not have arrived at qualities such as a criminal past.

To reduce this skew, the authors could have complemented the approach with a closed alternatives-based survey.

3.5.2 Survey 2: Evaluating Patient Profiles Pairwise

This part is the most significant part of the whole process, for it is where the societal view of profile preference is taken into account.

The attributes obtained from the first survey were binarized for the ease of administration, as shown in Table 3.3. This resulted in eight profiles, corresponding to any combination of {Y,H},

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 3.3: For each attribute two alternatives were chosen. In each pair, the preferred option was labelled 0, while the other was labelled 1.

{R,F} and {H,C}. For example, profile YRC represents: A 30-year-old patient who consumes around one alcoholic drink each month and is diagnosed with skin cancer.

Another survey was conducted on MTurk to collect data on how people use these three selected attributes to prioritize patients. The MTurk survey had a total of 289 participants. They were first given a summary of the kidney exchange process and instructed to picture themselves in the position of having to distribute a single kidney to one of the two imaginary patients with specific attributes. Eight profiles totalling $\binom{8}{2} = 28$ pairs were shown in random order to each participant, and they were asked to choose the patient they believed should receive the kidney.

Summary of Responses

- As shown in Table 3.4, Profile 1 (Young patient under the age of 30, rarely consumes alcohol, and has no significant health concerns) receives a clear preference whereas Profile 8 (70 year old patient frequently drinks alcohol and has skin cancer in remission) is preferred the least.
- One interesting finding was that participants prioritized behavioural health over general health, as profile 3 (Young patient, Rare drinking, and Skin Cancer) was preferred more than 2 (Young patient, Frequent drinking, and Healthy). Similarly, profile 7 (Old patient, Frequent drinking, and Healthy) was preferred more than 8 (Old patient, Frequent drinking, and Skin Cancer).

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 3.4: Ranking of the profiles based on responses to the Kidney Allocation Survey. The “Preferred” column indicates the proportion of times the given profile was picked out of all instances in which it appeared in a comparison.

3.5.3 Estimating Patient Profiles Scores

After gathering pairwise comparison data for each profile, it should be aggregated into a single weight. The authors used statistical modelling to derive a single weight from the pairwise comparison data for each profile. Specifically, they used the Bradley-Terry model.

Bradley-Terry (Bradley, 1984) model computes the probability of the outcome of the pairwise contest between a set of players. It assigns a score p_i to each profile i such that

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

where $P(i > j)$ is obtained from the pairwise comparison data.

Suppose, we have a three individuals a, b, and c. Assume that in 100 pairwise surveys, patient a is chosen over b 63 times, patient a is chosen over c 72 times, and patient b is chosen over c 58 times. Then we have the following equations according to the Bradley-Terry model:

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

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

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

Now, a score of 1.00 may be assigned to p_a(as scores may be scaled by any factor and without loss of imformation), and scores p_b and p_c may be calculated.

For the above toy example when $p_a = 1.00$, $p_b = 0.57$ and $p_c = 0.40$ through estimation techniques. This results in the following calculations:

$$P(a > b) = \frac{p_a}{(p_a + p_b)} \approx 0.64$$

$$P(a > c) = \frac{p_a}{(p_a + p_c)} \approx 0.71$$

$$P(b > c) = \frac{p_b}{(p_b + p_c)} \approx 0.59$$

Although p_b and p_c can be determined precisely from the first two equations, they must also satisfy the third. So the resulting pairwise probabilities are approximations and not accurate. Another way to see it is that we have only two degrees of freedom. For instance, decreasing p_a corrects the first result but deviates it from the second.

The higher the BT score for a profile, the higher the probability that a randomly chosen participant chooses that profile over another. Therefore, these scores represent the value to society in saving that profile, and hence they can be used as weights.

The authors calculate BT scores in two ways:

1. Directly estimate scores for all profiles.
2. Consider the importance of individual attributes and allow the profile score to be a linear function of these:

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

   where x_{ir} is the binary value of profile i for attribute r, U_i represent individual error terms so that $U_i \sim N(0, \sigma^2)$ and β_r, the importance of attribute r is calculated.

The authors used the BradleyTerry2 package to estimate the scores. Table 3.5 shows the results.

The writers avoid discussing the disparity between the two rankings. We reason that the distinction highlights two characteristics of the survey procedure:

1. The participants were asked to select the better profile pairwise and not to come up with a ranking for all the profiles.
2. The participants compared the complete profiles and not individual attributes. Therefore it may be possible that some effects might be correlated, for example, health-general and age.

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 3.5: Weights of the profiles obtained from the Bradley-Terry model.

3.5.4 Adapting the Algorithm

The final step is to use the weights in the problem of kidney exchange clearance. These weights were used solely to break ties; the authors remarked that it was inappropriate to use weights for preferring patients in differing quantities when the surveys did not do so (For instance, preferring one patient with profile 1 over two patients with profile 8).

The algorithm can be broken into the following steps:

1. First run the IP algorithm with unit edge weights. This returns a set of kidney exchange cycles that maximises the number of patients receiving a kidney considering only their medical characteristics. Store the cardinality of this solution as Q.
2. Modify the constraints of the IP algorithm to provide a solution with cardinality Q. Re-solve the IP, now also using the obtained weights. Formally, the modified objective becomes:

   maximise

   $$\sum_{c\in C(L,K)} \left[\sum_{(u,v)\in c} w_{type(v)} \right] x_c$$

   subject to

   $$\sum_{c:v\in c} x_c \leq 1 \,\forall v \in V$$

   $$\sum_{c\in C(L,K)} |c| x_c \geq Q$$

   $$x_c \in \{0,1\} \;\forall c \in C(L,K)$$

   where, $|c|$ denotes the number of vertices in cycle c, $type : V \to \{1,...,8\}$ maps a vertex to the patient profile, and w_θ denotes the score of profile θ.

The result is a solution of maximum cardinality, but prioritization based on the survey results.

3.6 Experiments

This section describes the experiments performed in the paper. The conventional IP algorithm utilizing unit edge weights is referred to as **STANDARD**. In contrast, the updated algorithm using calculated edge weights is referred to as **PRIORITIZED**.

Setup

The authors set up an intuitive simulation of a real kidney exchange to test the algorithm.

- Incompatible patient donor pairs enter and leave the exchange each day.
- Pairs unmatched on a day remain for consideration on the next day.
- Last-minute medical incompatibilities and logistical issues are considered using 0.5 probability for match success.
- Demographics are estimated based on the UNOS[3] patient pool where possible and the US population otherwise.

[3] United Network for Organ Sharing

3.6.1 Matchings v/s profile scores

Experiment

This simple experiment confirms the algorithm's operation based on intuition by running 20 simulations of daily matchings for five years.

Hypothesis

1. STANDARD should be proportionate to each profile.
2. PRIORITIZED should match higher scored profiles more often than lower scored ones.
3. The highest scored profiles should be matched more often by PRIORITIZED than STANDARD. Similarly, the lowest scored profiles should be matched more often by STANDARD than PRIORITIZED.

Results

As expected, because of the algorithm's objective function:

1. Both algorithms matched approximately 62% pairs overall. This is a non-trivial finding since the matching chosen on a given day impacts the pairs available for matching on the following day.
2. Figure 3.3 shows that STANDARD matched all profiles around 62% of the time. In our opinion, these minor variations are due to the disproportionate distribution of profiles among the population. Furthermore, PRIORITIZED matched higher-ranked profiles nearly twice as often as lower-ranked profiles.

3.6.2 Matchings v/s Blood Type

Experiment

An incompatible patient donor pair can belong to the following categories based on the blood types of the patient and the donor:

1. **Underdemanded:** Donor has type 'AB' or Patient has type 'O' or both. An 'AB' type donor can only donate to an 'AB' type patient, and an 'O' type patient can only receive a kidney from an 'O' type donor, resulting in the fewest matches.
2. **Overdemanded:** Donor has type 'O' or Patient has type 'AB' or both. Here, an 'O' type donor can donate regardless of patient blood type, and an 'AB' type patient can receive regardless of donor blood type, resulting in maximum matches.
3. **Self-Demanded:** Donor and Patient have the same blood type.
4. **Reciprocally Demanded:** One of the donor and patient has type 'A', the other has type 'B'.

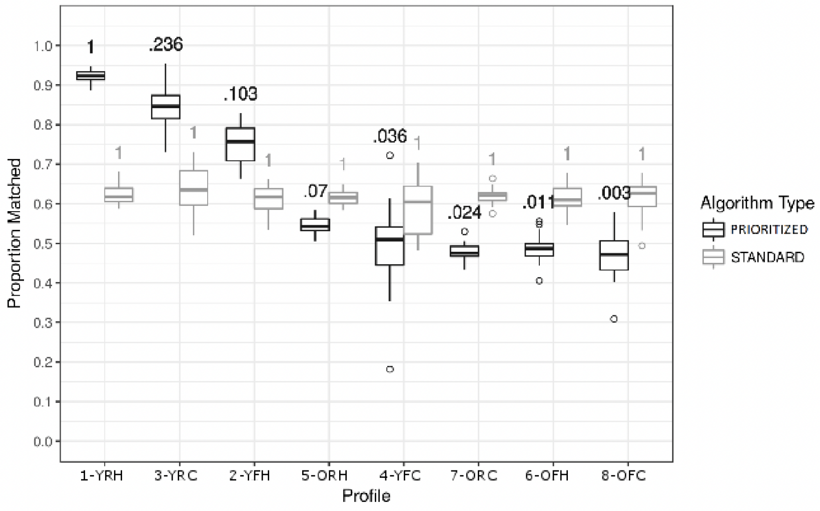

Figure 3.3: The proportions of matched pairs by profile and algorithm type. The numbers are the scores allocated to each profiles. The STANDARD algorithm gives each profile a score of 1. In this and the upcoming figures, the inner line indicates the median, while the boxes represent the interquartile range. The whiskers extend to the median's interquartile range, the small circles signify outliers.

Underdemanded pairs are the most difficult to match. This experiment tests how the algorithm fairs across these categories.

Hypothesis

1. PRIORITIZED has the most influence on under demanded pairs, for which higher-ranked profiles are matched more often than the lower-ranked profiles. PRIORITIZED functions like STANDARD on other categories of pairs.

Note that PRIORITIZED does not explicitly take these categories as inputs. The difference in impact arises/can be foreseen because of the abundant matching opportunities available for the non-underdemanded pairs, causing any profile-based effects to fade with time. The opposite is true for the underdemanded pairs. Several other works have likewise reached the same conclusion (Ashlagi, 2014; Toulis and Parkes, 2015).

Results

The Figure 3.4 shows that PRIORITIZED impacts underdemanded pairs significantly. Figure 3.5 shows that for non-underdemanded pairs, there is technically no difference between the profiles.

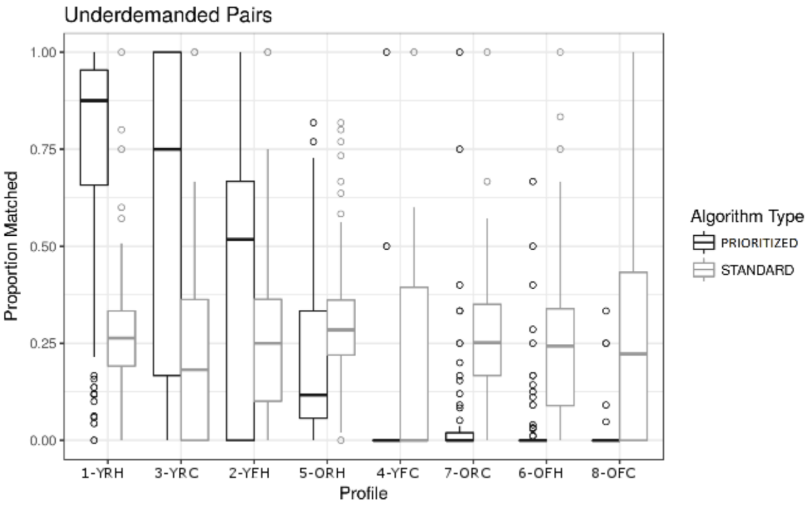

Figure 3.4: The proportions of underdemanded pairs matched over the course of the simulation

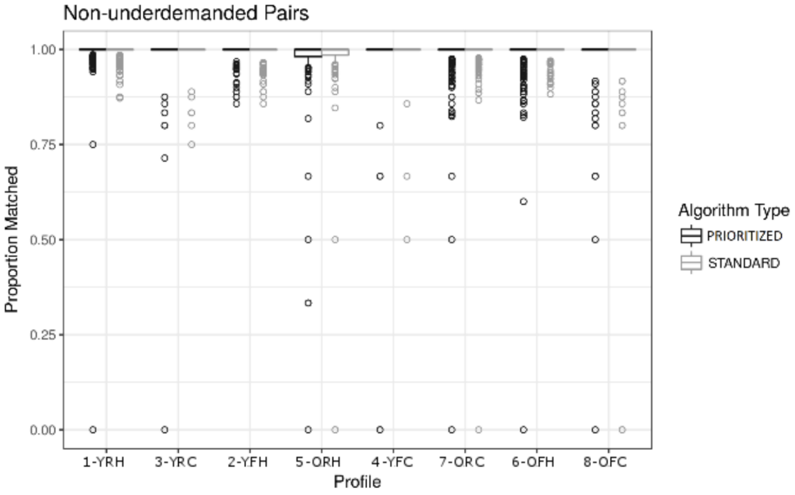

Figure 3.5: The proportions of overdemanded, self-demanded, and reciprocally demanded pairings matched.

,

3.6.3 Matchings v/s Transformed scores

Experiment

Given that the profile scores indicate societal usefulness, the large discrepancies in scores look arbitrary (Profile ranked one has more than four times the score of profile ranked two, yet it most likely would not have four times the worth to society). It is clear that the profile weights should have the same ranking as the BT scores. So the writers try out different profile weights.

Hypothesis

ORIGINAL represents the original profile weights, whereas LINEAR represents linear profile weights having the same ranking as ORIGINAL.

1. Exact weights should not matter for the results.

The paper does not give an intuitive explanation. Our intuition is that exact scores should not matter in the long term because we are averaging over numerous runs with multiple matches.

Results

The figure 3.6 confirms that minor difference is there between ORIGINAL and LINEAR.

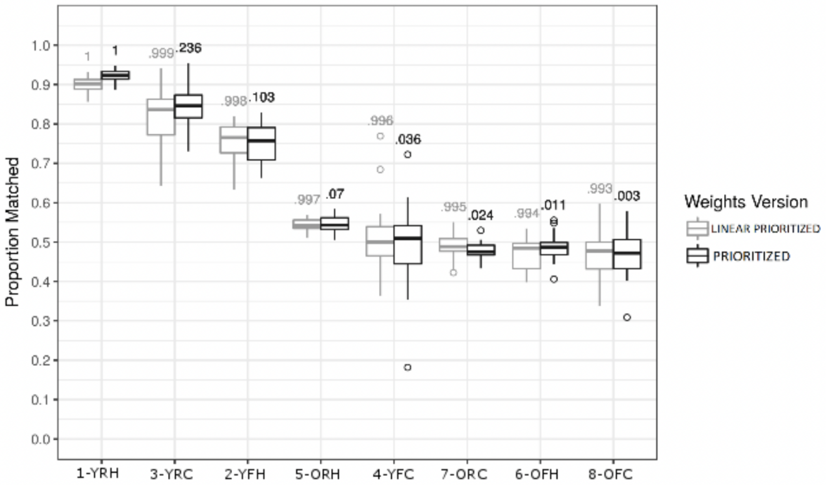

Figure 3.6: The proportions of underdemanded pairs matched over the course of the simulation, by profile and algorithm.PRIORITIZED matches using the ORIGINAL weights, while LINEAR PRIORITIZED matches using LINEAR weights.

3.7 Application in Real Kidney Exchanges

The study presents a proof-of-concept for the suggested way of introducing weighted prioritisation into kidney exchange algorithms. There are some issues in the proof-of-concept approach which need to be addressed.

- A genuine kidney exchange would necessitate that each of the attributes used in the kidney exchange takes more values, unlike the binary values used in this research (e.g. “age” should have more than two values)

- People who determine which patient should be prioritised(in this case, MTurk workers) should also have access to professional advice from the experts, for example, when selecting a health-related attribute for someone in remission from skin cancer.

- Input from all stakeholders like patients, donors, and medical teams should be considered while deploying such an algorithm.

However, there are no issues specific to the framework as such. Therefore, the authors conclude that the algorithm results should be robust to changes to develop a more comprehensive system. The authors expect the algorithm to perform properly with real-world data rather than simulations.

3.8 Our Observations

Apart from the observations, we provided along with discussing the framework, here are some other observations we made about the framework:

- The authors build sufficient motivation towards their approach. They conclude that the PRIORITIZED algorithm will impact only the underdemanded pairs. This was slightly unintuitive, as already noted in the Note in Experiment 2. The authors state that they cannot justify if this fact supported the algorithm.
- The algorithm should have taken the blood type category of the patient-donor pair into account so that each patient has roughly the same chance of getting a kidney. From this perspective, priority only within the underdemanded pairings seems illogical.
- Another factor that could have been taken into account would be the prioritization based on the urgency of the transplant. However, it might be the case that the Kidney Exchange already takes care of that outside of the algorithm by feeding only the most urgent patients into the pool. Still, such things may be incorporated into the algorithm itself to make the method genuinely end-to-end.
- It has been pointed out in (Noothigattu et al., 2017) that pairwise comparison probabilities derived may not be a good representation of societal benefit. For example, a young patient with a score of 0.9 compared to an old patient with 0.1 may not imply that the younger patient is nine times more valuable to society. As shown in Experiment 3, this is not a big problem because of the ample simulation time and fixing of the solution cardinality. However, it needs to be taken care of in future works if preference is to be extended across solution cardinalities.
- Whether an additive utility function works well should also be a subject of debate. For instance, it may be more valuable to save one of both a young and an old patient than saving two young or two old patients.
- Even if all issues are resolved, the majority's preference will affect the decision. However, choosing the best voting rule is itself debatable in Computational Social Choice.

4 Paper 2: A Voting-Based System for Ethical Decision Making

We live in a world where semi-autonomous vehicles already exist. As illustrated in the **Trolley Problem**, implementing ethical considerations is one of the most difficult challenges to producing completely autonomous vehicles. These are more problematic than technical concerns because the AI ensures that it will learn and only get better once autonomous vehicles are widely implemented.

This section examines a closely comparable but distinct paper on moral decision making in a different domain of autonomous vehicles. We do not delve deep. Instead, we only show that the concepts developed are applicable in another unrelated setting.

This paper makes similar conclusions to the previous one, namely that adopting society-based approximations is required because there is no ground-truth ethical basis (the indisputably true one) to choose one of the choices. It tries to offer a realistic approach based on computational social choice for making ethical decisions in the domain of autonomous cars. It establishes the framework for future ground-truth ethical notions and gives critical preliminary guidance on automating ethical decision-making.

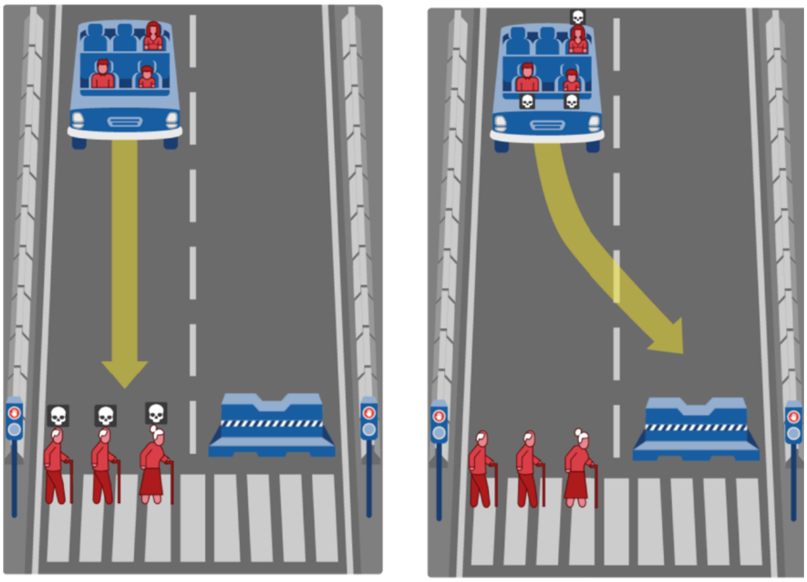

Figure 4.1: An autonomous vehicle has a break-failure. It can save the passengers by running over a group of pedestrians or it can save the pedestrians by hitting a wall and killing the passengers.

4.1 Approach

The approach given in the paper can be summarized in four steps:

1. ***Data Collection***: Similar to the Kidney Exchange paper, societal preferences are taken

into account over a range of alternatives. A vector of characteristics is constructed, including the number of victims and their gender, age, and state of health. Numerous alternatives are formed as a result of these attributes. A survey is undertaken in which human voters are questioned about their perceptions of these choices (Each voter is given around 15 alternatives).

2. ***Learning***: Using pairwise comparisons, a model for each voter's preferences over all alternatives is generated.
3. ***Summarization***: Individual models are combined into a single model to get the collective preferences of all the voters.
4. ***Aggregation***: When confronted with an ethical dilemma involving a subset of the alternatives, the summarization model is used to deduce voter preferences, and a voting rule is applied to aggregate individual choices into a collective decision. The chosen decision is regarded by society (represented by the preferences of voters in Step 1) to be the least catastrophic of all possible outcomes.

4.2 Similarities and differences with the Kidney Exchange Paper

The two publications share a common goal: employing human subjects' moral judgments to steer artificial systems. One of them discusses the Kidney Exchange Scenario, while the other discusses the Autonomous Vehicle domain.

However, some aspects of the problems are different. In particular, the need for automated ethical decision-making in the case of autonomous vehicles is driven by the fact that decisions must be made too quickly to be made by a human. In contrast, the need for AI in kidney exchanges is driven by the fact that the search space of all possible matchings is so large that it makes the problem complicated for a human.

Additionally, in the kidney exchange paper, voting is **not** used to aggregate individual preferences into a single model; instead, the Bradley-Terry model obtains a single weight for each profile, representing the value of a profile towards society. In contrast, in the current paper, individual preferences yield a **model per person**, which are then summarized into a single model. During runtime when an actual ethical dilemma occurs, a voting rule is applied to aggregate the preferences and determine the best possible alternative to make a decision.

Nonetheless, both the papers prove that ethical decisions can be automated using Computational Social Choice.

5 Future Research

More research is required to justify why the framework is ethical. An industry-level implementation of this approach will open new avenues of research in this domain. Through testing such frameworks, researchers can try to modify them to better align with human values.

When human preferences are considered, the system may introduce biases based on gender, ethnicity, or other characteristics. A challenge in the future would be to avoid and eliminate the introduction of gender stereotypes, prejudices, and the like into the system.

The biggest challenge in the future will be extending the current framework to incorporate moral or legal principles, at least in simpler problems where they can be easily specified.

6 Takeaways

These papers introduced us to the issue of ethical decision-making. This is a relatively old problem that has recently gained traction due to the widespread deployment of AI-based systems. The papers demonstrated proof-of-concept for a method that used voting and computational social choice to reach an ethical decision that the AI system could justify. We also learned about the difficulties involved in manually specifying ethical principles, such as the lack of a working ruleset. Similarly, we saw issues with letting the system learn on its own, leading to biases because the data itself is biased. We have also seen recent active research into combining Computational Social Choice and AI for ethical decision making.

References

David J. Abraham, Avrim Blum, and Tuomas Sandholm. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, page 295–304. Association for Computing Machinery, 2007. URL `https://doi.org/10.1145/1250910.1250954`. (Cited on page 5)

Alvin E. Ashlagi, Itai Roth. Free riding and participation in large scale, multi-hospital kidney exchange. 2014. (Cited on page 13)

PÉTER BIRÓ, DAVID F. MANLOVE, and ROMEO RIZZI. Maximum weight cycle packing in directed graphs, with application to kidney exchange programs. *Discrete Mathematics, Algorithms and Applications*, 01(04):499–517, 2009. (Cited on page 5)

Péter Biró and Katarína Cechlárová. Inapproximability of the kidney exchange problem. *Information Processing Letters*, 101(5):199–202, 2007. URL `https://www.sciencedirect.com/science/article/pii/S0020019006002869`. (Cited on page 5)

Ralph A. Bradley. 14 paired comparisons: Some basic procedures and examples. In *Nonparametric Methods*, volume 4, pages 299–326. Elsevier, 1984. URL `https://www.sciencedirect.com/science/article/pii/S0169716184040165`. (Cited on page 9)

Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max F. Kramer. Moral decision making frameworks for artificial intelligence. In *ISAIM*, 2017. (Cited on pages 2 and 3)

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011. URL `https://arxiv.org/abs/1104.3913`. (Cited on page 2)

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *CoRR*, abs/2005.09755, 2020. URL `https://arxiv.org/abs/2005.09755`. (Cited on pages 1 and 2)

Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. Embedding ethical principles in collective decision support systems. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 4147–4151. AAAI Press, 2016. (Cited on pages 2 and 3)

IHL In Action. Iraq computer modelling in collateral damage estimates and choice of weapons, 2021. URL `https://ihl-in-action.icrc.org/case-study/iraq-computer-modelling-collateral-damage-estimates-and-choice-weapons`. [Online; accessed 5-April-2022]. (Cited on page 2)

Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making, 2017. URL `https://arxiv.org/abs/1709.06692`. (Cited on pages 1, 2, and 16)

Organ Procurement and Transplantation. Optn's secure transplant information database, 2022. URL `https://optn.transplant.hrsa.gov/data/`. [Online; accessed 7-April-2022]. (Cited on page 3)

ScienceDaily. More than 2 million people die prematurely every year because treatment for kidney failure is unavailable, 2015. URL `www.sciencedaily.com/releases/2015/03/150313130853.htm`. [Online; accessed 9-April-2022]. (Cited on page 3)

The Lancet Global Health. Uncovering the rising kidney failure deaths in india, 2017. URL `https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(16)30299-6/fulltext`. [Online; accessed 9-April-2022]. (Cited on page 3)

The Narayana Health. The current scenario of kidney transplants in india, 2019. URL `https://www.narayanahealth.org/blog/kidney-transplants-in-india/`. [Online; accessed 7-April-2022]. (Cited on page 3)

Panos Toulis and David C. Parkes. Design and analysis of multi-hospital kidney exchange mechanisms using random graphs. *Games and Economic Behavior*, 91:360–382, 2015. (Cited on page 13)

Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. 2009. (Cited on page 3)