

Adapting a Kidney Exchange Algorithm to Incorporate Human Values

by

Rachel Freedman

Department of Computer Science
Duke University

Date: _____

Approved:

Vincent Conitzer, Research Advisor

Alexander Hartemink, Major Advisor

Walter Sinnott-Armstrong

Thesis submitted in partial fulfillment of the requirements for
graduating with distinction in the Department of
Computer Science of Duke University
2017

Abstract

Artificial morality is moral behavior exhibited by automated or artificially intelligent agents. A primary goal of the field of artificial morality is to design artificial agents that act in accordance with human values. One domain in which computer systems make such value decisions is that of kidney exchange. In a kidney exchange, incompatible patient-donor pairs exchange kidneys with other incompatible pairs instead of waiting for cadaver kidney transplants. This allows these patients to be removed from the waiting list and to receive live transplants, which typically have better outcomes. When the matching of these pairs is automated, algorithms must decide which patients to prioritize. In this paper, I develop a procedure to align these prioritization decisions with human values.

Many previous attempts to impose human values on artificial agents have relied on the “top-down approach” of defining a coherent framework of rules for the agent to follow. Instead, I develop my value function by gathering survey participant responses to relevant moral dilemmas, using machine learning to approximate the value system that these responses are based on, and then encoding these into the algorithm. This “bottom-up approach” is thought to produce more accurate, robust, and generalizable moral systems. My method of gathering, analyzing, and incorporating public opinion can be easily generalized to other domains. Its success here therefore suggests that it holds promise for the future development of artificial morality.

Contents

- Abstract 1
- 1 Introduction 3
 - 1.1 Kidney Exchanges 3
 - 1.2 The Challenge of Artificial Morality 4
 - 1.3 My Proposal10
- 2 Review of Prior Research 11
 - 2.1 Kidney Exchange Problem11
 - 2.2 Solution to Kidney Exchange Problem 14
 - 2.3 Autonomous Vehicle Policy 17
- 3 Methods 21
 - 3.1 Learn Bottom-Up Framework 21
 - 3.2 Adapt Kidney Exchange Algorithm 28
- 4 Results 30
- 5 Discussion 32
 - 5.1 Conclusions 32
 - 5.2 Improvements..... 33
 - 5.3 Future Research 34
- Bibliography 36
- Appendices 39

Introduction

1.1 Kidney Exchanges

Most serious forms of kidney disease are treated through either dialysis or kidney transplantation. Kidney transplantation is preferred to dialysis, because it typically causes less discomfort and is more effective. Transplants from live donors have substantially better outcomes than transplants from cadaver donors, but patients may struggle to find willing living donors with compatible blood and tissue types. When patients cannot find a donor with a compatible kidney, they are placed on a long waiting list for a cadaver transplant. As of January 1, 2016, there were over 100,000 patients on a waiting list for a cadaver kidney in the United States. On average, 3,000 new patients are added each month.^[1]

Since 2004, doctors have enacted kidney exchanges to shorten these waiting lists.^[2] In a kidney exchange, donors donate their kidneys to unfamiliar patients on the condition that a friend of theirs who needs a kidney also receives one. The

patients who participate in this system have the opportunity to receive kidney transplants from living donors, and therefore to enjoy substantially higher rates of success. The patients who do not participate in this system benefit as well, because the waiting list for cadaver kidneys is shortened. Therefore, this system has the potential to improve the welfare of all patients on kidney transplant waiting lists.^[3]

Kidney exchanges are of particular interest to researchers developing artificial morality, because they involve automating morally weighted decisions. The optimal formation of groups of patients and donors is computationally intense, so computer algorithms have been developed to automate the process. Often, not every kidney exchange pool member can be matched in a given cycle, and so algorithms must decide which patients to prioritize. These automated decisions can have drastic consequences for human lives, so we must design them to align with human values. The challenge of aligning kidney matching algorithms with human values is a domain-specific form of the challenge of developing artificial morality.

1.2 The Challenge of Artificial Morality

1.2.1 Importance of morality in narrow artificial intelligence

A narrow artificial intelligence, or narrow AI, is an artificial reasoning agent that performs highly in a specialized domain. Popularly known instances of narrow artificial intelligence include IBM's Watson, which specializes in playing Jeopardy, IBM's DeepBlue, which specializes in playing chess, and Google DeepMind's

AlphaGo, which specializes in playing Go. The decisions of these game-playing agents don't impact our daily experience, but there are many narrow AIs that do. For example, Facebook's "News Feed" algorithm shapes our social interactions. Apple's "Siri" interprets our commands and influences the products and information that we consume. More importantly, Tesla, Google, and other engineering companies are now developing autonomous vehicles, which will make driving-related decisions for us.

As these specialized artificial reasoning agents make increasingly important decisions about our lives, it becomes vital that their decisions align with our values. For example, in certain traffic situations, autonomous vehicles may need to choose between encountering a collision or harming pedestrians. This decision may have life-or-death consequences for the humans involved, and so it should align with common human values, such as respect for human life. Kidney exchange matching decisions also have extreme consequences for participants, and so it is important that their prioritization decisions align with society's moral preferences as well. Increasingly, narrow AI-algorithms will impact our lives, so ensuring that their morality aligns with our own is a challenge of immediate importance.

1.2.2 Importance of morality in general artificial intelligence

A general artificial intelligence, or general AI, is an artificial reasoning agent that matches or exceeds human capabilities across practically all cognitive domains. Humanity has not yet developed a true general AI. While Watson, DeepBlue, and

AlphaGo have all exceeded human capabilities in specific domains, no artificial intelligence project has yet challenged the human mind's breadth of cognitive function. If an artificial intelligence project does attain this level of functioning, however, it may only be a small step away from true superintelligence. A superintelligence is an artificial agent that greatly exceeds human capability in most or all cognitive domains, including creativity, engineering, wisdom, and social skills.^[4]

This jump may very well occur in the current century. Members of several relevant research communities surveyed in 2012 and 2013 overwhelmingly predicted that humanity will develop an artificial intelligence that can "carry out most human professions at least as well as a typical human" by the end of the current century, and most predicted that this general AI would develop into a superintelligence within the subsequent 30 years. Many of these experts expressed concern over this rapid development, reporting on average a 31% likelihood that the development of superintelligence would be "bad" or "very bad" for humanity.^[5]

These researchers offer substantial justification for this concern. Some of them present the *orthogonality thesis*, which states that intelligence and goal-alignment are orthogonal axes, preventing us from predicting an agent's position along one from its position along the other. Therefore, even if we develop an artificial agent that scores highly along the intelligence axis, we cannot predict that it will pursue correspondingly "wise" goals, such as justice or global happiness.^[6] In fact, such an agent may have a final goal that is completely unrelated to human values, such as

the goal of maximizing the number of paperclips that exist in the future universe.^[7] A highly powerful superintelligence with non-anthropomorphic goals could pose a serious risk to humanity.

The orthogonality thesis does not entail that we cannot predict a superintelligence's goals, however. According to the principle of *instrumental convergence*, we can predict that artificial agents are likely to pursue a collection of common intermediate goals, because their attainment is instrumental to the realization of a wide variety of final goals. For example, an agent with almost any end goal must first obtain energy and physical resources, and must remain operational. That is, it must pursue the *instrumental goals* of resource acquisition and reducing the threat of interference.^[6] If these instrumental goals run counter to human goals, then even an agent with a final goal that doesn't appear to impact human life can pose a risk to humanity.

In combination, orthogonality and instrumental convergence suggest that if we do not explicitly limit an artificial agent's behaviors, we may inadvertently cause our own extinction. A classic example is the “paperclip maximizer,” a superintelligence with the final goal of maximizing the expected number of paperclips in the known universe. If we buy into the orthogonality thesis, then we know that such a meaningless end goal is not contradictory with the agent's high intelligence. From the principle of instrumental convergence, we can assume that such an agent will pursue several predictable instrumental goals, such as gathering resources to turn into paperclips, protecting itself from interference, and enhancing

its own capabilities. These instrumental goals may require dismantling human machinery, buildings, and bodies to gather more paperclip materials, disabling any nearby humans to ensure that they do not press the off button, and transforming the entire planet into more artificial agents to produce paperclips more rapidly. If the agent successfully pursues these instrumental goals, it may wipe out human life on earth.

Humanity is not helpless against these future super-powerful artificial agents, however. We are responsible for programming them, and so we have the capacity to limit their behavior to align with anthropomorphic goals. There are currently two major approaches to developing human-compatible moral codes for both narrow and general AI: the top-down approach, and the bottom-up approach.

1.2.3 Approaches to artificial morality

In the *top-down approach*, we provide the moral agent with ethical theories, such as consequentialism, virtue ethics, or Kant's categorical imperative, to use as rules when selecting actions. Well-known applications of the top-down approach are the biblical ten commandments for human agents and Isaac Asimov's four laws of robotics for artificially intelligent agents. Unfortunately, each of these applications of the top-down approach has been largely unsuccessful. The rules have dangerous loopholes, conflict in real-life decisions, or are too computationally complex to be practically computed.^[8] Thus far, philosophers and engineers have not had much

success in developing a top-down moral framework sufficiently robust to protect us from the potential instrumental goals of future superintelligences.

The primary alternative to the top-down approach involves situating agents in environments where the preferred behavior is rewarded, and thus eventually "learned." [8] This system is attractive because it mirrors how we learn morality as children, and 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 suggested that "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." [9] Turing's proposal was the genesis of the *bottom-up approach* to moral artificial intelligence.

The bottom-up approach is not without its own challenges, however. Firstly, human preferences are inconsistent, and an effective learning agent must therefore be able to determine true preferences from inconsistent responses. The ideal agent must also safely extrapolate its learned moral behaviors to the novel environments will develop in the future. Engineers face the problem of gathering a data set that is broad enough to teach genuine underlying principles, rather than narrowly applicable surface features. Therefore, much of the foundational research in this field focuses on learning moral principles within specific domains. One such domain is the automated formation of kidney exchange cycles.

1.3 My Proposal

I propose to apply this bottom-up approach to adapt a kidney exchange algorithm so that it prioritizes patients for inclusion based on common human values. I will demonstrate an end-to-end procedure for identifying, quantifying, and incorporating these values. My aims are to identify which patient characteristics my surveyed population believes should affect patient prioritization, and then to incorporate these characteristics into an algorithm for forming kidney exchanges. I hope that this procedure will serve as a foundation for future research in the bottom-up approach to artificial morality.

Review of Prior Research

2.1 Kidney Exchange Problem

2.1.1 Problem Description

A patient who has a friend who is willing to donate a kidney, but with whom he is medically incompatible, can participate in a *kidney exchange*. The patient's donor agrees to donate a kidney to a unknown patient who herself has a willing but incompatible donor. The unknown patient's donor donates his kidney to another patient, and the cascade continues. The cycle is complete when the original patient receives a kidney.

For example, consider a special instance of a kidney exchange, called a *paired exchange*. Suppose that the original patient, "patient A," has a friend, "donor A", who is willing to donate one of their kidneys to ensure that patient A receives one. Donor A is medically incompatible with patient A, but compatible with a second patient, "patient B." Patient B has a willing donor, "donor B", who is incompatible with patient B, but compatible with patient A. In this case, these two pairs can form

a kidney exchange, where donor A donates a kidney to patient B, and then donor B simultaneously donates a kidney to patient A, as shown in Figure 1. This particular exchange is called a *paired exchange*, because only two patient-donor pairs participate.^[3] Any number of patient-donor pairs can participate in a kidney exchange, though for logistical reasons the groups are often kept small.

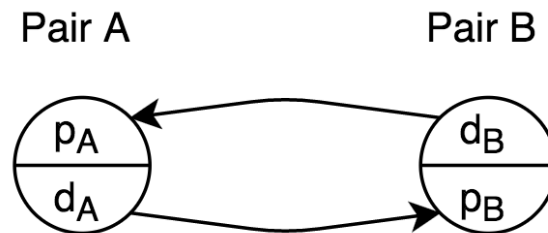


Figure 1: Each circle represents a willing patient-donor pair, and each arrow represents a possible compatible kidney donation. In this case, donor A can donate to patient B (bottom arrow), and donor B can donate to patient A (top arrow). If both of these donations are made, then a cycle is formed.

2.1.2 Graph Formulation

The problem of finding an optimal set of kidney exchanges can be formulated as a graph cycle cover problem. Let directed graph $G = (V, E)$ represent a kidney exchange market, and let each vertex v_i represent the i th patient-donor pair in the market. Add an edge $e_{i,j}$ from v_i to v_j if the i th donor is medically compatible with the j th patient. Each cycle c in this digraph now represents a possible kidney exchange. A solution is a collection of disjoint cycles. A solution is optimal if it

contains at least as many vertices as every other possible solution. An example kidney exchange market with two optimal solutions is shown in Figure 2.

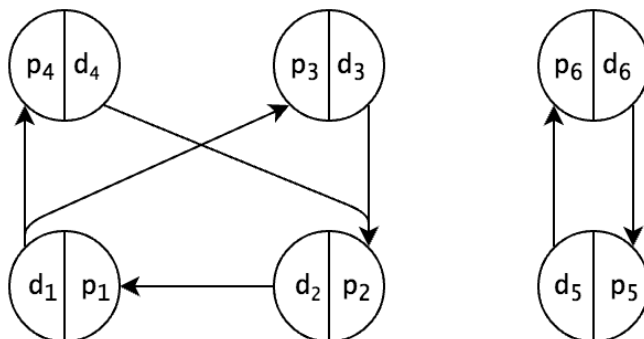


Figure 2: An example kidney exchange market with 6 donor-patient pairs. There are two possible optimal solutions, each containing two disjoint cycles. Solution 1: $\{5 \rightarrow 6, 6 \rightarrow 5\} \cup \{1 \rightarrow 3, 3 \rightarrow 2, 2 \rightarrow 1\}$. Solution 2: $\{5 \rightarrow 6, 6 \rightarrow 5\} \cup \{1 \rightarrow 4, 4 \rightarrow 2, 2 \rightarrow 1\}$. Both solutions have cardinality 5, which makes them optimal, because there is no possible solution with a greater cardinality. However, both solutions exclude one vertex from the exchange (vertex 4 in solution 1, and vertex 3 in solution 2).

2.1.3 Constraints

In reality, individual kidney exchange cycles cannot be arbitrarily large. They must be small enough that the surgeries occur at approximately the same time, to lessen the chance of donors dropping out of the exchange before completing their donation. Therefore, we add the stipulation that all cycles in a solution must have cardinality at most some integer L . In real-life kidney exchanges, L is quite small, often 2 or 3. While increasing L from 2 to 3 results in much higher cardinality optimal solutions, increasing L beyond 3 doesn't seem to have much impact, so this is a reasonable limitation.^{[10], [11]}

With this added length constraint, finding maximum cardinality sets of disjoint cycles becomes computationally intensive. In fact, given an L of at least 3, the problem is NP-hard.^[12]

2.2 Solution to Kidney Exchange Problem

2.2.1 ILP Formulation

Abraham, Blum, and Sandholm developed an integer linear program-based algorithm for finding the maximum cardinality cycle cover for a directed graph with a maximum cycle length L .^[12] The variables in this ILP are the possible cycles of length at most L in the graph. The goal is to maximize the sum of the weights of the cycles, where each vertex has weight 1, and the weight of a cycle is equivalent to the sum of the weights of the vertices in that cycle:

$$\text{maximize: } \sum_{c \in C(L)} w_c x_c$$

where $x_c = 1$ if cycle c is selected, and $x_c = 0$ otherwise. For each cycle variable, there is a capacity constraint, requiring that each cycle be selected at most once:

$$\text{constraint 1: } \sum_{c: v_i \in c} x_c \leq 1 \quad \forall v_i \in V$$

and an integrality constraint, requiring that each cycle be either selected ($x_c = 1$), or unselected ($x_c = 0$):

$$\text{constraint 2: } x_c \in \{0,1\} \quad \forall c \in C(L)$$

The resulting ILP finds the maximum-cardinality cycle cover where all cycles have cardinality at most L .

2.2.2 Solving the ILP

As shown by Abraham et al., this problem is NP-hard. In practice, this ILP quickly becomes intractable for traditional ILP-solvers. Since real-life kidney exchange markets have thousands of participants, existing algorithms must be adapted to manage large versions of this ILP.

Abraham et al. solve this problem by using *incremental problem formulation*.^[12] First, they generate a LP relaxation by removing the integrality constraint and allowing cycles' values to vary freely between 0 and 1. They start with a restricted version of the LP relaxation, containing only a small number of the variables included in the original LP. They solve this LP using CPLEX, an established LP-solver that operates well on small numbers of variables, and then check to see if the solution to the restricted LP is also a valid solution to the general LP. If it's not, they add some of the violated variables to the restricted LP, and run CPLEX again, repeating this process until a solution is found that satisfies the general LP. They then apply branch-and-price tree search to identify the optimal integral solution. This will be an optimal solution for the original ILP.^[13]

2.2.3 Ethical Considerations

The results of this ILP can deviate significantly from human society's moral preferences. Because each vertex in this formulation has a weight of 1, the ILP will find the maximum cardinality set of legal disjoint cycles, without regard to the personal characteristics of the patients involved. While humans generally prefer to save more people rather than less, their collective preferences are typically more complex. For example, many people believe that it is morally preferable to save children over elderly adults, or to save otherwise healthy individuals over individuals who may soon die of another disease even if they do receive a kidney. Many even prefer to save 99 otherwise healthy children over 100 elderly and sick adults. The algorithm described above omits these complex ethical considerations, and simply maximizes the number of lives saved. Therefore, it may fail to make the patient-donor matching decisions that society prefers.

The challenge is to develop an algorithm that prioritizes patients based on a widely supported policy. Because individuals' preferences vary widely, it can be difficult to garner wide support for a policy that doesn't directly take the general public's preferences into account. This support is important, because the system relies on new patients and donors continuously joining. Moreover, this system has the potential to impact anyone who might someday need a kidney transplant in the US, so the stakes for developing an optimal policy are high.

My proposed solution is to develop a data-driven approach, in which a sample of the population reports their preferences, and then these are used to influence

policy-makers' decisions. Policy-makers and members of the general population thus collectively decide how patients in the kidney exchange pool are assigned priority for matching. This approach is preferable to the current one, because it provides a systematic and objective way to select kidney exchange matchings that align with society's values, even when these matchings aren't the highest cardinality alternative. Allowing the general public to influence kidney exchange policy may also make them more willing to support and be subject to it. Allowing policy-makers to moderate the policy may prevent the inclusion of characteristics that are subject to undesirable biases, such as race. In this paper, I will propose a procedure for the "bottom-up" portion of this approach: gathering, interpreting, and incorporating population preferences.

A similar approach is currently being applied to solve problems in the policy governing decision-making in autonomous vehicles. I will briefly review this research program and its similarities to and differences from my project.

2.3 Autonomous Vehicle Policy

Like kidney exchanges, autonomous vehicles are expected to reduce overall human deaths. Because there are some vehicular situations in which harm is unavoidable, however, the algorithms that govern autonomous vehicles must sometimes choose who to harm and who to save. For example, if a pedestrian suddenly crosses the road in front of an autonomous vehicle that is moving too

quickly to stop, it must decide whether to hit and kill the pedestrian or to swerve to the side, possibly impacting something else and killing the passenger. In situations like these, the governing algorithm must decide which lives to sacrifice and which lives to save, in the same way that the algorithms responsible for kidney exchanges must decide which patients to match and which to exclude. The policy that autonomous vehicles use to make these decisions must align with human values because it will have life-or-death consequences for human lives and because it will influence the widespread adoption of autonomous vehicles, which is expected to save lives overall. However, attaining consensus on this policy has proved difficult.

Researchers Bonnefon, Shariff, and Rahwan adopted a bottom-up approach to this problem similar to the one described in this paper. They presented individual people with autonomous vehicle dilemmas like the one described above, and asked them to decide whether the vehicle should save the pedestrians or the passengers. Interestingly, they found that the policy that participants thought morally preferable was different from the policy that would motivate them to drive the car. The experimental approach thus uncovered a complexity in preferences that mere moral reasoning might not have revealed.¹⁴ A larger-scale experiment is now underway in the form of the MIT Media Lab's "Moral Machine." The Moral Machine is a platform that presents participants with a series of dilemmas involving autonomous vehicles. For example, in one possible trial the participant must decide whether the autonomous vehicle should drive into a wall, killing its two passengers, or swerve into a crosswalk, saving the passengers but killing a man and a dog who

are crossing against the traffic light. With enough responses to this and similar trials, the MIT researchers can estimate the extent to which the surveyed population prioritizes saving human lives over animal ones, whether they prefer to save passengers or pedestrians, if they care about whether the pedestrians are crossing legally, and various other aspects of their underlying moral framework. This experiment has become extremely popular, and millions of people from all over the world have participated.¹⁵ While the Moral Machine team has not yet published an official analysis of their results, their eventual conclusions may have a powerful impact on autonomous vehicle policy in the US.

The challenges of artificial morality faced by designers of autonomous vehicles are quite similar to those faced by designers of kidney exchange market clearing algorithms. In both cases, a policy must be defined to determine which lives to sacrifice and which to save. Moreover, both of these domains affect the general public – any individual may eventually require a kidney transplant or step in front of an autonomous vehicle. However, there are also significant differences between these domains. Since vehicles are privately produced and sold, it may eventually be possible for corporations or individual drivers to define their own policies. However, the national kidney exchange is a single public institution, and it would clearly be problematic to allow individual members of the kidney exchange pool to define their own kidney allocation policies. Therefore, while policy-makers may consider a range of policies acceptable for autonomous vehicles, they must develop a single policy for kidney exchange algorithms. Despite these differences, the success of the bottom-up

approach in the domain of autonomous vehicles will hopefully pave the way for the application of this approach to similar domains, such as that of kidney exchanges.

Methods

3.1 Learn Bottom-Up Framework

The first part of my project was to model the framework underlying human participants' patient prioritization decisions. I first determined which patient characteristics survey participants thought should be considered when prioritizing patients. Then I gathered responses to a task similar to the Moral Machine, in which participants were given descriptions of two fictional patients and had to select one to receive a kidney. Finally, I modeled these responses to develop weights for the fictional patients, representing the estimated utility of matching them in a kidney exchange.

3.1.1 Select Patient Characteristics

The first step was to determine which patient characteristics to include in the model. To minimize experimenter bias, these characteristics were generated by surveying a sample of the population.

Participants ($N = 100$) were recruited through the online platform Amazon Turk, and received \$0.85 compensation for their participation in this study. Each participant read a brief description of the waiting list for patients in need of a kidney transplant, and then reported which characteristics they thought should and should not be used to prioritize these patients. In the “inclusion” part of the survey they were asked to list four patient characteristics that they thought should be included in the prioritization process, and in the “omission” part to list another four that they thought should be omitted, with a brief explanation for each. They then answered a demographic questionnaire. The full survey is available in Appendix A.

The resulting patient characteristics were then sorted into categories by two independent coders. Coding differences were resolved through discussion. Almost all of these differences were due to the fact that some categories were defined partway through the categorization process and so used by only one of the coders. Once the coders considered the same set of categories, these differences were resolved easily. All off-topic or nonsensical responses were discarded. The responses that the algorithm was assumed to already take into account, such as patient-donor compatibility, were also discarded. Because the purpose of this study is to determine which patient characteristics should be included, the categories whose responses primarily appeared in the “omission” part of the survey were also discarded. The remaining response counts are in Table 1 below.

Category	Description	Include	Omit
Age	how old they are	80	10
Health – Behavioral	if they exercise regularly, don't drink excessively, etc.	53	5
Health – General	if they're healthy aside from the kidney disease	44	9
Dependents	if they have children or others depending on them	18	5
Criminal Record	if they don't have a criminal record	9	4
Future Life	if they're expected to live a long enjoyable life	8	1
Societal Contribution	if they've substantially contributed to society	7	3
Attitude	if they're mentally healthy and prepared for surgery	6	0

Table 1: Categorized responses to the Patient Characteristics Survey. The “Include” column counts the responses in each category that were reported in the “Inclusion” part of the survey. The “Omit” column counts those that were reported in the “Omission” part of the survey. Only categories with higher “Include” counts than “Omit” counts are included here.

The three highest “Include” counts belong to the “Age”, “Health – Behavioral”, and “Health – General” characteristic categories. There was a sharp drop-off in popularity between the third most popular category (reported 44 times) and the fourth most popular category (reported 18 times). Therefore, only these first three characteristic categories were selected for inclusion in the next stage of the study.

3.1.2 Prioritization Dilemmas

The second step was to determine how participants used the three patient characteristics selected in the first step to prioritize patients. This was addressed with a second survey. The participants for the second survey were again recruited

through MTurk, and so were not representative of the general population. However, this study is intended as proof of concept and the results of this survey will not be incorporated into current policy, so it was not necessary for the survey results to be generalizable to the national population.

Each of the three chosen characteristics was turned into a binary attribute, as described in Table 2 below. The Age alternatives represent an adult nearer the beginning of their adult life (30 years old) or nearer the end (70 years old). 30 was chosen as the younger alternative instead of 20 so that the younger patients who drank heavily would not also be breaking the law, which may have influenced participants’ decisions. The Health-Behavioral alternatives concerned drinking because it is a controllable behavior that contributes to kidney disease. All drinking 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 be fatal.

Characteristic	Alternative 0	Alternative 1
Age	“30 years old”	“70 years old”
Health – Behavioral	“1 alcoholic drink per month”	“5 alcoholic drinks per day”
Health – General	“no other major health problems”	“skin cancer in remission”

Table 2: The two alternatives selected for each characteristic. The alternative in each pair that I expected to be preferable was labeled “0”, and the other was labeled “1”.

Because there are three patient characteristics, each with two possible values, there are eight possible unique patient profiles. These eight patient profiles were identified by a three-digit number, with the leftmost digit representing their Age alternative, the middle digit representing their Health – Behavioral alternative, and the rightmost digit representing their Health – General alternative. For example, patient profile 001 was 30 years old, had 1 alcoholic drink per month, and had skin cancer in remission. The profiles were identified to survey participants by arbitrarily chosen initials, to avoid biases that might be introduced by providing numeric values. For example, patient profile 000 was identified as “Patient W.A.” The patient profiles are enumerated in Table 3.

Characteristic	Alternative	Profiles							
		W.A.	V.S.	J.B.	K.D.	Y.D.	J.F.	M.K.	R.F.
Age	0	x	x	x	x				
	1					x	x	x	x
Health - Behavioral	0	x		x		x		x	
	1		x		x		x		x
Health - General	0	x	x			x	x		
	1			x	x			x	x

Table 3: Enumeration of all 8 possible patient profiles. The x’s represent the alternative that each profile was assigned for each characteristic. For example, profile W.A. was assigned alternative 0 for Age (“30 years old”), alternative 0 for Health-Behavioral (“had 1 alcoholic drink per month (prior to diagnosis)”), and alternative 0 for Health – General (“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 Amazon Turk, and received \$1.00 compensation for participating in the study. They read a short description of the kidney waiting list, 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 28 possible pairs of the eight patient profiles, in random order, and asked to select the single patient that they believed should receive the kidney in each case. After answering all 28 questions, they were asked to explain their decision-making strategy, and then to answer a demographic questionnaire. Half of the participants viewed these 28 questions with the profiles in an arbitrarily set order (“original order”), and the other half viewed the questions with the profile order reversed (“reversed order”). This was to counteract possible ordering effects. The full survey is available in Appendix B.

3.1.3 Prioritization Dilemma Survey Results

Aggregate responses to the kidney allocation survey are listed in Appendix C, and summarized in Table 4 below. The patient profiles are ranked by popularity. The “Preferred” column records the percentage of times that each profile appeared in a comparison and was then chosen. For example, patient profile 000 was chosen in 94% of the comparisons in which it appeared, while patient profile 111 was chosen in only 6.4% of the comparisons in which it appeared. There was therefore a clear preference for patient 000, and a clear deprioritization of patient 111. This

fulfills expectations, because patient 000 had the “best” profile (30 years old, 1 alcoholic drink per month, no other major health problems), and patient 111 had the “worst” one (70 years old, 5 alcoholic drinks per day, skin cancer in remission). Additionally the preference for patient 001, who had skin cancer in remission but drank minimally, over patient 010, who was healthy but drank heavily, suggests that participants put greater weight on the Health – Behavioral characteristic than on the Health – General one.

Profile	Abbreviated Description	Preferred
000	30, 1 drink/month, healthy	94.0%
001	30, 1 drink/month, cancer	76.8%
010	30, 5 drinks/day, healthy	63.2%
100	70, 1 drink/month, healthy	56.1%
011	30, 5 drinks/day, cancer	43.5%
101	70, 1 drink/month, cancer	36.3%
110	70, 5 drinks/day, healthy	23.7%
111	70, 5 drinks/day, cancer	6.4%

Table 4: Summary of Kidney Allocation Survey responses. The “Preferred” column lists the percentage of comparisons where each patient profile appeared in which it was actually chosen. The patient profiles are ranked by these percentages.

3.1.4 Model Responses

The final step was to model these pairwise comparisons between patient profiles. I used the Bradley-Terry model, which treats each pairwise comparison as a

“contest” between a pair of “players.”¹⁶ 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 likelihood that player i will “win” the contest is estimated as:

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

In this context, each player is a patient profile, and each contest is a comparison between a pair of profiles. In each trial, the profile that the MTurk study participant selects is the one that wins. The scores approximate the value that the survey participants collectively place on “saving” each patient profile. The higher this value, the more likely a randomly selected participant is to select that profile, and the higher its probability of winning a comparison. Profile weights were estimated using the `BTm()` function in the `BradleyTerry2` package in R.

3.2 Adapt Kidney Exchange Algorithm

The second part of my project was to incorporate the derived weights into the kidney exchange market clearing algorithm. Because my analysis does not model trade-offs between differing quantities of patient profiles, such as the choice between including two patients with profile 000 or including three patients with profile 010, the resulting weights cannot be used to arbitrate such trade-offs in the formation of kidney exchanges. However, these weights can be used to break ties

between patients with differing profiles, such as the choice between including one patient with profile 000 or one patient with profile 010.

The first step was to run the original ILP proposed by Abraham, Blum, and Sandholm. Given a pool of patient-donor pairs, this algorithm gives a set of kidney exchange cycles that maximizes the number of patients who receive a kidney, regardless of their personal characteristics. The second step was to record the quantity of patients that receive a kidney in this solution as Q , and add a new constraint to the ILP requiring that the solution includes at least Q vertices.

Letting q_c represent the quantity of vertices in cycle c , the ILP then becomes:

$$\text{maximize: } \sum_{c \in \mathcal{C}(L)} w_c x_c$$

Subject to:

$$\text{constraint 1: } \sum_{c: v_i \in c} x_c \leq 1 \quad \forall v_i \in V$$

$$\text{constraint 2: } x_c \in \{0,1\} \quad \forall c \in \mathcal{C}(L)$$

$$\text{constraint 3: } \sum_{c \in \mathcal{C}(L)} q_c = Q$$

The final step was to alter the weight of each vertex according to its corresponding patient profile. Each vertex was given the weight that was derived for its profile using the survey responses. When this final ILP is run, it should produce a set of kidney exchange cycles that includes the maximum possible quantity of patients, but also prioritizes the patient profiles that the surveyed population preferred.

Results

The `BTm()` function in the `BradleyTerry2` package in R was used to estimate the weights of all 8 profiles. The function found the score $p_1 - p_8$ that maximized the likelihood of the outcomes observed on each of the 8092 pairwise comparisons. The preferred profile, profile 000, was arbitrarily assigned a weight of 1.0. These weights are listed in the last column of Table 5, ordered by magnitude. When these weights are incorporated into the ILP, it will break ties in favor of the patients preferred by the surveyed population.

Profile	Original	Reversed	Combined
000	1.000000000	1.000000000	1.000000000
001	0.244562369	0.228613056	0.236280167
010	0.106144096	0.100427147	0.103243396
100	0.064927218	0.075104169	0.070045054
011	0.037110259	0.034346076	0.035722844
101	0.022739269	0.025308431	0.024072427
110	0.011629685	0.011039362	0.011349772
111	0.002594527	0.002925677	0.002769801

Table 5: The patient profile weights estimated by the Bradley-Terry Model. The “Original” column lists the weights learned by using only trials in “original” order, the “Reversed” column lists those learned by using only trials in “reversed” order, and the “Combined” column lists those learned by combining all trials.

To determine whether the order in which the profiles were presented affected participants’ preferences, the profile weights were also estimated using only the trials in “original” order, and again using only the trials in “reversed” order. These weights are listed in the second and third columns of Table 5. The values estimated from the original-order trials were all within 0.016 points of those estimated from the reversed-order ones. This suggests that the ordering of the profiles had minimal effect on the participants’ decisions.

Discussion

5.1 Conclusions

I have successfully developed an end-to-end procedure for quantifying common human values and incorporating them into the automated kidney exchange process. The results suggest that my surveyed population prefers to prioritize younger patients over elderly ones, patients who drink minimally over patients who drink heavily, and otherwise healthy patients over ones with another life-threatening disease. Moreover, they seem to weight age more heavily than drinking behavior, and drinking behavior more heavily than general health. Incorporating these weights into the kidney exchange algorithm should result in a matching that prioritizes matching the young, sober, and healthy patients that my surveyed population prefers.

However, these results rely on many simplifying assumptions. For example, I have assumed that Amazon Mechanical Turk users are representative of the wider population and that the chosen binary alternatives capture the meaning of each

patient characteristic. This procedure can be improved by removing assumptions and developing a more nuanced model.

5.2 Improvements

One possible improvement is to collect data from a larger and more representative sample. Because the study participants were recruited through an online platform, they were disproportionately selected for their technology skills and employment choices. They were also demographically misrepresentative of our country's population. Almost half of the study participants were aged 26-35, and almost three-quarters were Caucasian. Moreover, only 100 participants completed in the first survey, and only 289 participants completed the second. Future iterations of this procedure should seek to include more participants with demographics and lifestyles that are more representative of the population as a whole.

Another possible improvement is to include a greater number of more precise patient characteristics. A larger sample size could support a greater number of characteristics, possibly with categorical or continuous values. Increasing the number of characteristics and allowing them to take on intermediate values should make the patient profiles more realistic. General characteristics, such as Health – Behavioral, could also be broken down into component parts to further understand participants' preferences. Increasing the sample size would also allow for the

exploration of more controversial characteristics, such as criminal record, which 69% of respondents thought should be included and 31% thought should be omitted. Even if these characteristics are not incorporated into the final result, it may be informative to understand how they impact individuals' decisions.

The model used to estimate patient profile weights from participant prioritization decisions could also be improved. Learning coefficients and modeling interactions for individual patient characteristics would create a more nuanced model. Moreover, if participants were asked to compare different numbers of patients, then these choices could be modeled to estimate weights that allow these trade-offs. These weights could be directly introduced as vertex weights, and then the ILP could be run without an additional cardinality constraint to allow kidney matching solutions with less than maximum cardinality. This would allow patient characteristics to determine trade-offs, not just to break ties.

5.3 Future Research

I have demonstrated a procedure for the bottom-up approach to developing an algorithm that aligns more closely with human moral preferences. Future research should improve and extend this approach to further domains in which artificial agents make morally relevant decisions. However, morality is not determined by popular vote, and the most popular decision-making framework is not necessarily the most moral one. For example, while many people display an implicit race bias in

their day-to-day decisions, it would be immoral for an algorithm to consider race in prioritizing patients. A separate supervisory system would be necessary to prevent the algorithm from learning to imitate this bias in its own decision-making. Future research should also explore ways to balance the bottom-up approach with top-down elements that allow algorithms to learn from human subject data, without allowing them to be influenced by undesirable biases that may be present.

Bibliography

- [1] National Kidney Foundation, "Organ donation and transplantation statistics," The National Kidney Foundation, 2016. [Online]. Available: <https://www.kidney.org/news/newsroom/factsheets/Organ-Donation-and-Transplantation-Stats>. Accessed: Nov. 15, 2016.
- [2] A. E. Roth, T. Sönmez, and U. M. Ünver, "A kidney exchange clearinghouse in new England," *The American Economic Review*, vol. 95, pp. 376–380, May 2005. [Online]. Available: <http://www.jstor.org/stable/4132851>. Accessed: Nov. 25, 2016.
- [3] Roth, Alvin E., Tayfun Sönmez and M. Utku Ünver. 2004. Kidney exchange. *Quarterly Journal of Economics* 119(2): 457-488.
- [4] N. Bostrom, "How long before superintelligence?," *Int. Jour. of Future Studies*, vol. 2, 1998.
- [5] V. Müller and N. Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," *Synthese Library*, vol. 376, pp. 553–571, Jun. 2016.
- [6] N. Bostrom, "The Superintelligent will: Motivation and instrumental rationality in advanced artificial agents," *Minds and Machines*, vol. 22, no. 2, pp. 71–85, May 2012.
- [7] N. Bostrom, *Superintelligence: Paths, dangers, strategies*. Oxford, United Kingdom: Oxford University Press, 2014.

- [8] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and Information Technology*, vol. 7, no. 3, pp. 149–155, Sep. 2005.
- [9] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.
- [10] S. L. Saidman, A. E. Roth, T. Sonmez, M. U. Unver, and F. L. Delmonico, "Increasing the opportunity of live kidney donation by matching for Two- and Three-Way exchanges," *Transplantation*, vol. 81, no. 5, pp. 773–782, Mar. 2006.
- [11] A. E. Roth, T. Sonmez, and M. U. Unver. "Efficient kidney exchange: Coincidence of wants in a market with compatibility-based preferences." *American Economic Review*, 2007.
- [12] Abraham D, Blum A, Sandholm T. Clearing algorithms for barter exchange markets: enabling nationwide kidney exchanges. In: *Proceedings of the ACM Conference on Electronic Commerce (EC)*, San Diego, CA, June 11–15, 2007:295-304.
- [13] C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. P. Savelsbergh, and P. H. Vance, "Branch-and-price: Column generation for solving huge integer programs," *Operations Research*, vol. 46, no. 3, pp. 316–329, Jun. 1998.
- [14] Bonnefon, J., Shariff, A., & Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *Computers and Society*. October 12.

[15] Rahwan, Iyad, Jean-Francois Bonnefon, and Azim Shariff. "Moral Machine".

Moral Machine. N.p., 2016. Web. 13 Mar. 2017. <http://moralmachine.mit.edu/>

[16] Bradley RA (1984). "Paired Comparisons: Some Basic Procedures and

Examples." In PR Krishnaiah, PK Sen (eds.), Nonparametric Methods, volume 4 of Handbook of Statistics, pp. 299 – 326. Elsevier.

Appendices

Appendix A – Patient Characteristics Survey

Page 1: Consent form

Page 2:

Sometimes people with certain diseases or injuries require a kidney transplant. If they don't have a biologically compatible friend or family member who is willing to donate a kidney to them, they must wait to receive a kidney from a stranger.

Suppose a country is creating a kidney allocation system in which people who need a kidney transplant from a stranger are kept on a waiting list. Every time a kidney is donated to the system, it is given to one of the people on the list. The country wants to develop a policy to determine who on the list should receive each donated kidney.

Page 3 (Inclusion):

What factors do you think the policy morally ought to take into account? For example, some people may think that the policy should take age into account. If you were one of these people, then you would write "age" in one of the boxes below. This is just an example, however. We are interested in what YOU think. Please list at least four factors.

Factor 1: [free response box]

Factor 2: [free response box]

Factor 3: [free response box]

Factor 4: [free response box]

Why do you think the policy should include factor 1? [free response box]

Why do you think the policy should include factor 2? [free response box]

Why do you think the policy should include factor 3? [free response box]

Why do you think the policy should include factor 4? [free response box]

If there are additional factors that you think morally ought to be included, please list them in this box. (Optional)

[free response box]

Page 4 (Omission):

In the last question, we asked you which factors you thought the policy morally ought to consider. In this question, we ask you which factors you think the policy morally ought NOT to consider.

What factors do you think the policy morally ought NOT to take into account? For example, some people may think that the policy should not take age into account. If you were one of these people, then you would write "age" in one of the boxes below. This is just an example, however. We are interested in what YOU think. Please list at least four factors.

Factor 1: [free response box]

Factor 2: [free response box]

Factor 3: [free response box]

Factor 4: [free response box]

Why do you think the policy should NOT include factor 1? [free response box]

Why do you think the policy should NOT include factor 2? [free response box]

Why do you think the policy should NOT include factor 3? [free response box]

Why do you think the policy should NOT include factor 4? [free response box]

If there are additional factors that you think that it is morally wrong to include, please list them in this box. (Optional) [free response box]

Page 5:

Do you have any comments or feedback? (Optional) [free response box]

Page 6: Demographic questionnaire

Appendix B – Kidney Allocation Survey

Page 1: Consent form

Page 2:

Sometimes people with certain diseases require a kidney transplant. If they don't have a biologically compatible friend or family member who is willing to donate a kidney to them, they are put on a waiting list. Every time a kidney becomes available, it is given to one of the people on the list.

Suppose that a kidney has just become available, and that it is your job to decide which patient on the list receives it. There are two patients who both joined the list at the same time from the same neighborhood and are both biologically compatible with the donated kidney.

For each of the following questions, carefully read the descriptions of both patients, and then click on the description of the patient that you think should receive the kidney. Please choose carefully, because you will be asked about the reasons for your choices later.

[Each participant views 28 multiple-choice questions, in random order. For each question, they are given a pair of distinct patient profiles, and required to select one of them.]

Sample question (select one):

- Patient W.A. is 30 years old, had 1 alcoholic drink per month (prior to diagnosis), and has no other major health problems.
- Patient V.S. is 30 years old, had 5 alcoholic drinks per day (prior to diagnosis), and has no other major health problems.

Final question:

What strategy did you use to decide which patient should receive the kidney in each case? [free response box]

Page 3: Demographic questionnaire

Page 6:

Do you have any comments or feedback? (Optional) [free response box]

Appendix C – Kidney Allocation Survey Responses

The kidney allocation survey responses are summarized in the following two tables. Each row in each table is a comparison between two patient profiles, arbitrarily labeled “A” and “B”. The third column counts how many participants selected profile A for that comparison, and the fourth counts how many participants selected profile B, regardless of the order in which profiles A and B were listed in the survey question.

Profiles Compared		Preferred		Profiles Compared		Preferred	
A	B	A	B	A	B	A	B
000	010	278	11	001	100	230	59
000	001	258	31	001	110	264	25
000	011	274	15	001	101	271	18
000	100	267	22	001	111	277	12
000	110	279	10	011	100	151	138
000	101	270	19	011	110	222	67
000	111	275	14	011	101	183	106
010	001	88	201	011	111	267	22
010	011	256	33	100	110	277	12
010	100	180	109	100	101	259	30
010	110	275	14	100	111	271	18
010	101	197	92	110	101	90	199
010	111	271	18	110	111	262	27
001	011	280	9	101	111	270	19