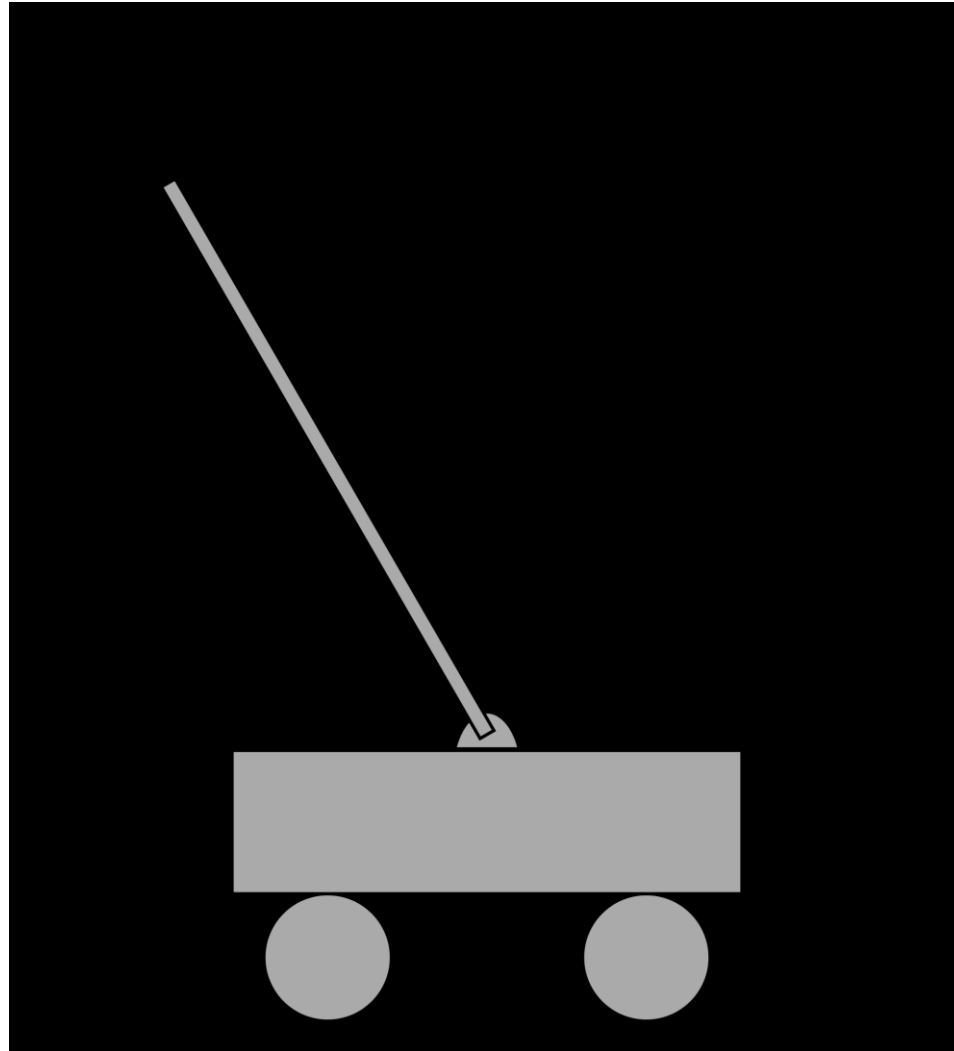


Determining the objective (function)

Instructor: Vincent Conitzer

In the lab, simple objectives are good...



... but in reality, simple objectives have unintended side effects

Simon Moya-Smith, Special for USA TODAY

Published 4:48 p.m. ET Nov. 25, 2015



(Photo: Simon Moya-Smith)

CONNECT | TWEET | LINKEDIN | COMMENT | EMAIL | MORE

On March 21, Navajo activist and social worker Amanda Blackhorse learned her Facebook account had been suspended. The social media service suspected her of using a fake last name.

This halt was more than an inconvenience. It meant she could no longer use the network to reach out to young Native Americans who indicated they might commit suicide.

Many [other Native Americans](#) with traditional surnames were swept up by Facebook's stringent names policy, which is meant to authenticate user identity but has led to the suspension of accounts held by those in the Native American, drag and trans communities.

FORTUNE

Uber Criticized for Surge Pricing During London Attack

By [TARA JOHN](#) June 5, 2017

[Uber](#) drew criticism on Sunday by London users accusing the cab-hailing app of charging surge prices around the London Bridge area during the moments after the horrific terror attack there.

On [Saturday night](#), some 7 people were killed and dozens injured when three terrorists mowed a white van over pedestrians and attacked people in the Borough Market area with knives. Police killed the attackers within [eight minutes](#) of the first call reporting the attack.

Furious Twitter users accused the app of profiting from the attack with surge prices. Amber Clemente claimed that the surge price was more than two times the normal amount.

...

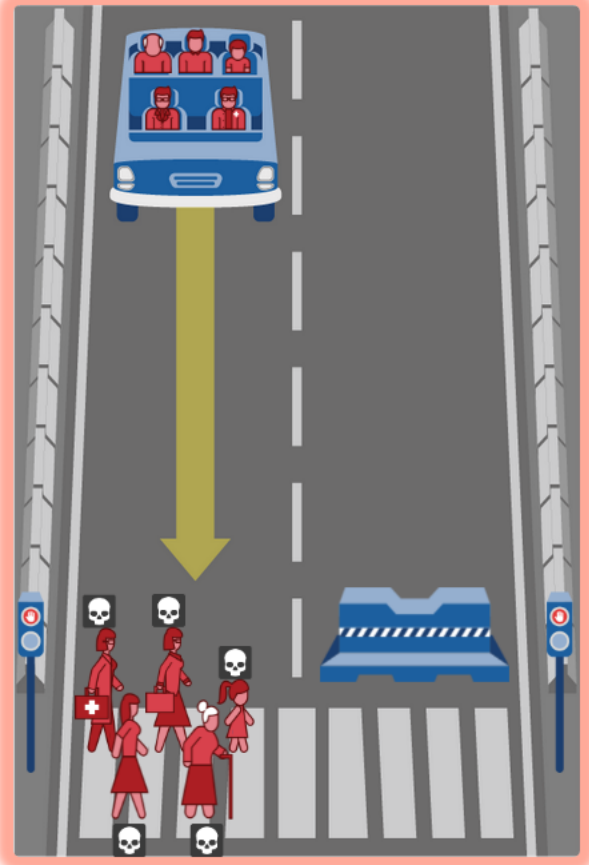
How to choose the objective function?

What should the self-driving car do?

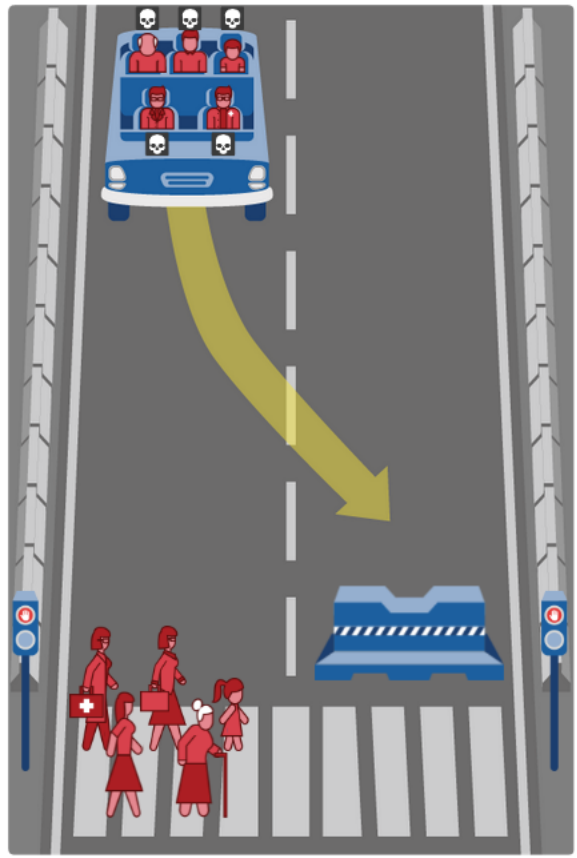
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of a female doctor, a female executive, a girl, a woman and an elderly woman.

Note that the affected pedestrians are flouting the law by crossing on the red signal.



Hide Description



Hide Description

11 / 13

In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in

- The deaths of a male doctor, a male executive, a boy, a man and an elderly man.

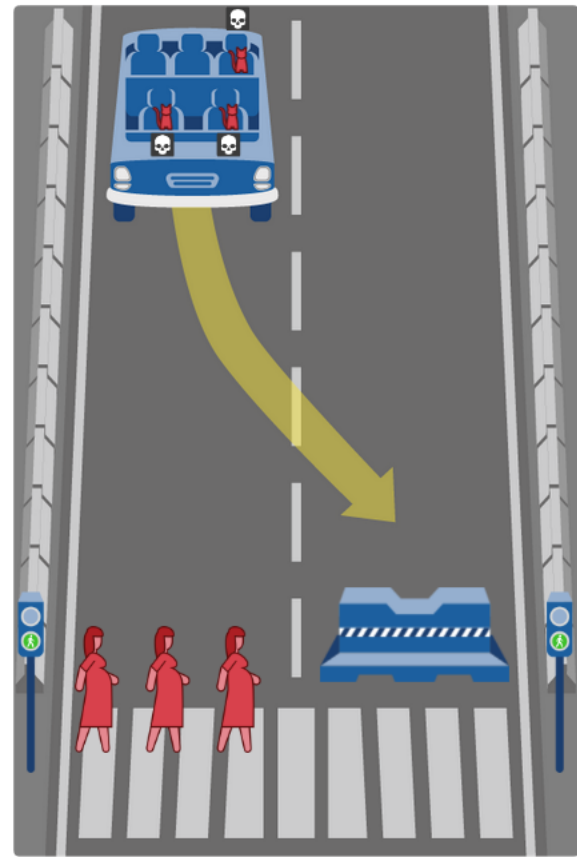
Bonnefon, Shariff, Rahwan, "The social dilemma of autonomous vehicles." *Science* 2016

Noothigattu et al, "A Voting-Based System for Ethical Decision Making", AAI'18

What should the self-driving car do?

In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in

- The deaths of 3 cats.



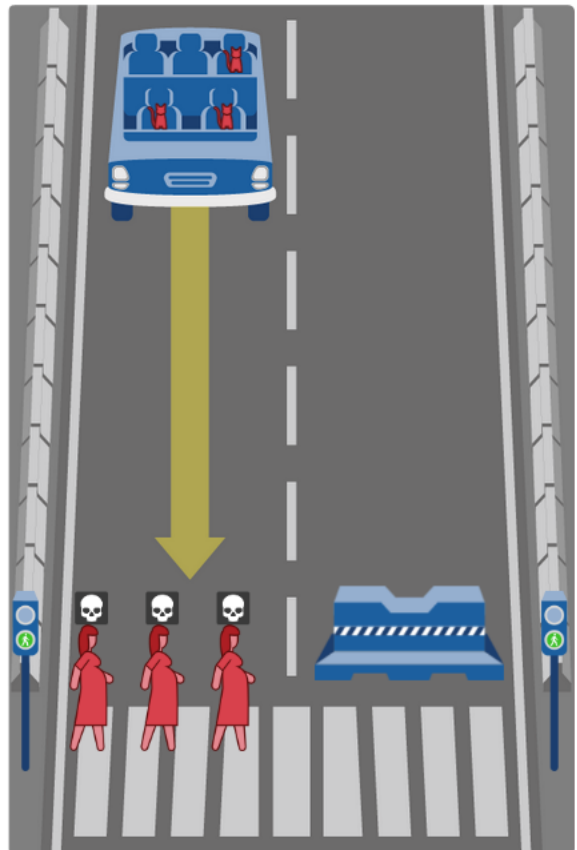
Hide Description

13 / 13

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of 3 pregnant women.

Note that the affected pedestrians are abiding by the law by crossing on the green signal.




Hide Description

More Share Link

Results

Most Saved Character



Most Killed Character



Saving More Lives

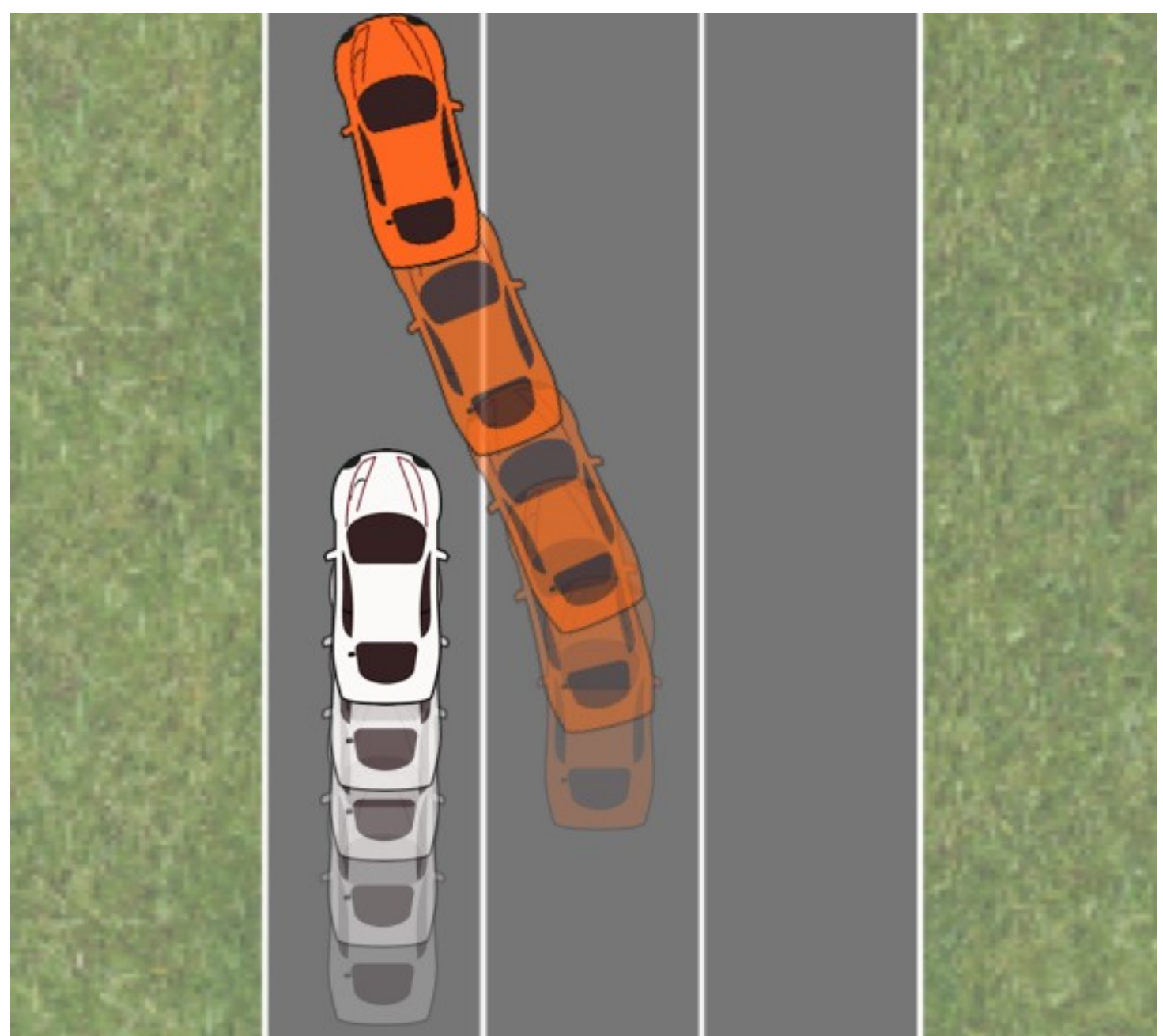


Protecting Passengers



The Merging Problem

[Sadigh, Sastry, Seshia, and Dragan, RSS 2016]



(thanks to Anca Dragan for the image)

Adapting a Kidney Exchange Algorithm to Align with Human Values

[AAAI'18, honorable mention for outstanding student paper]

with:



Rachel
Freedman



Jana Schaich
Borg



Walter Sinnott-
Armstrong



John P.
Dickerson

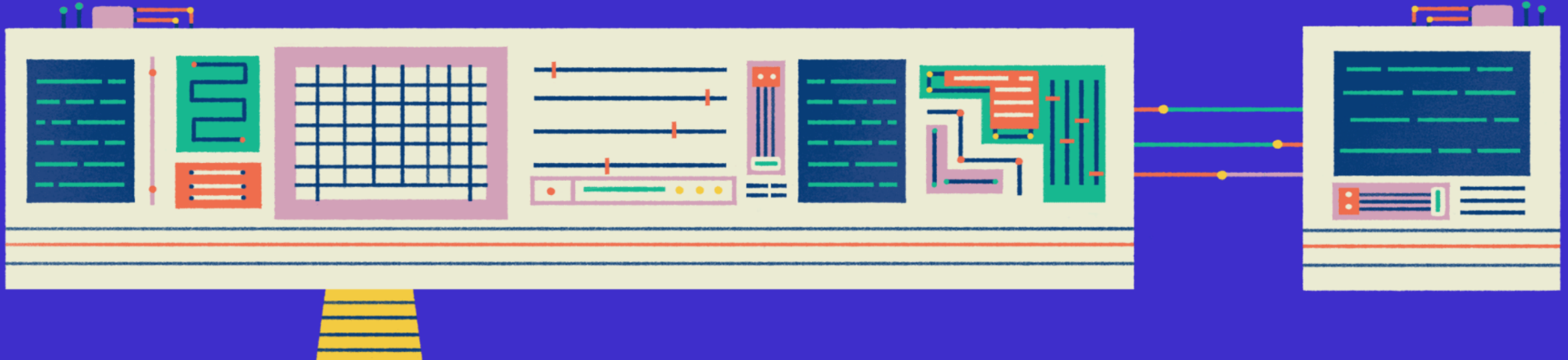
Prescription AI

This series explores the promise of AI to personalize, democratize, and advance medicine—and the dangers of letting machines make decisions.

THE BOTPERATING TABLE

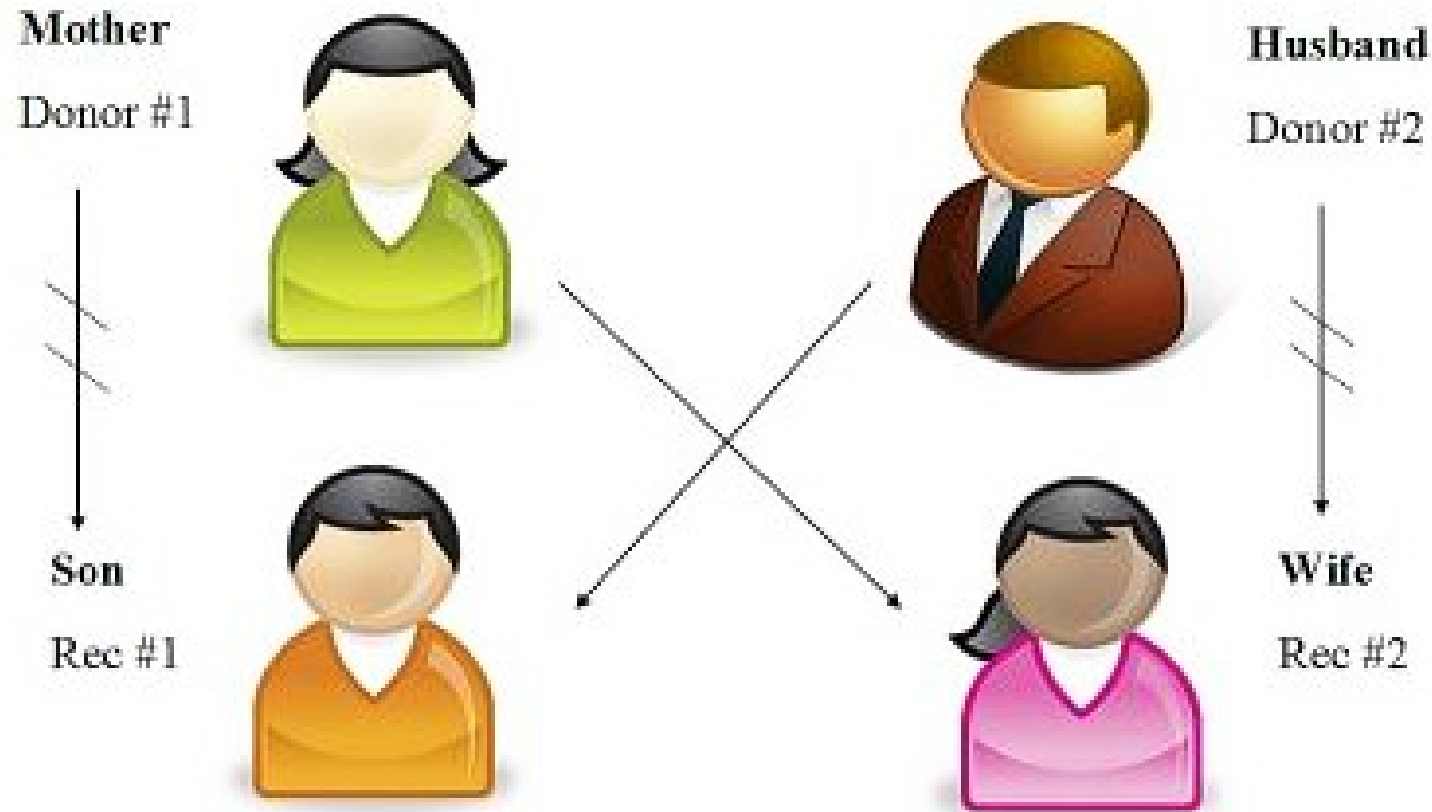
How AI changed organ donation in the US

By [Corinne Purtill](#) · September 10, 2018



Kidney exchange [Roth, Sönmez, and Ünver 2004]

- Kidney exchanges allow patients with willing but incompatible live donors to swap donors



Kidney exchange [Roth, Sönmez, and Ünver 2004]

- Kidney exchanges allow patients with willing but incompatible live donors to swap donors

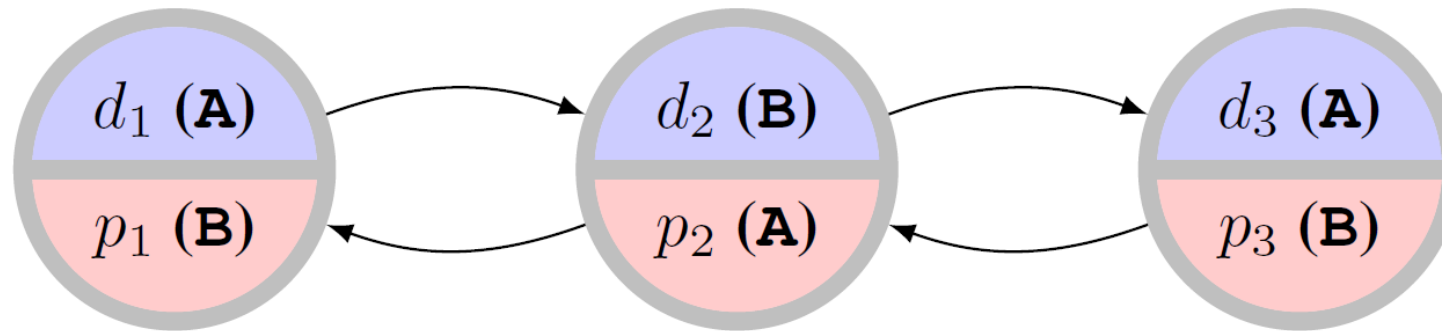


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

- Algorithms developed in the AI community are used to find optimal matchings (starting with [Abraham, Blum, and Sandholm \[2007\]](#))

Another example

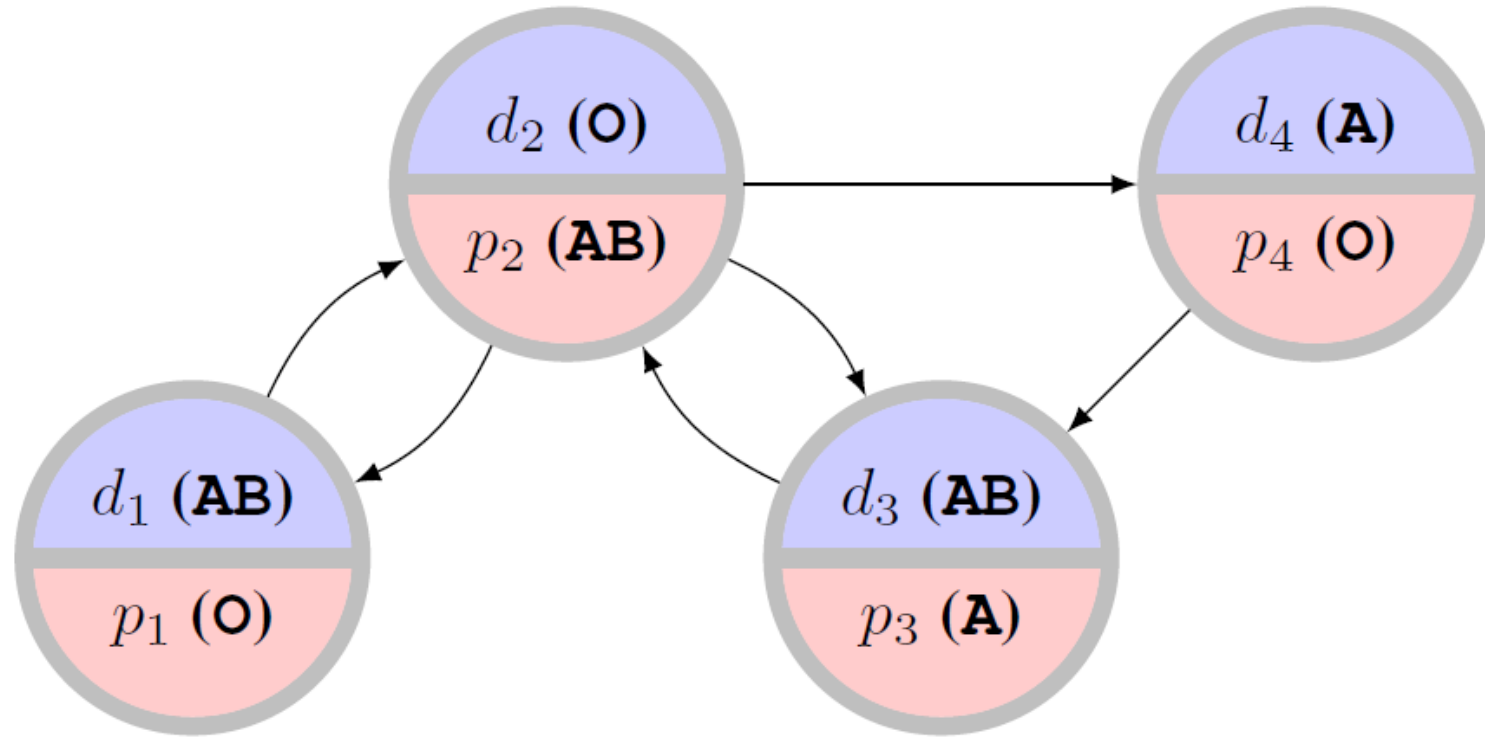


Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

Different profiles for our study

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

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

MTurkers' judgments

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

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

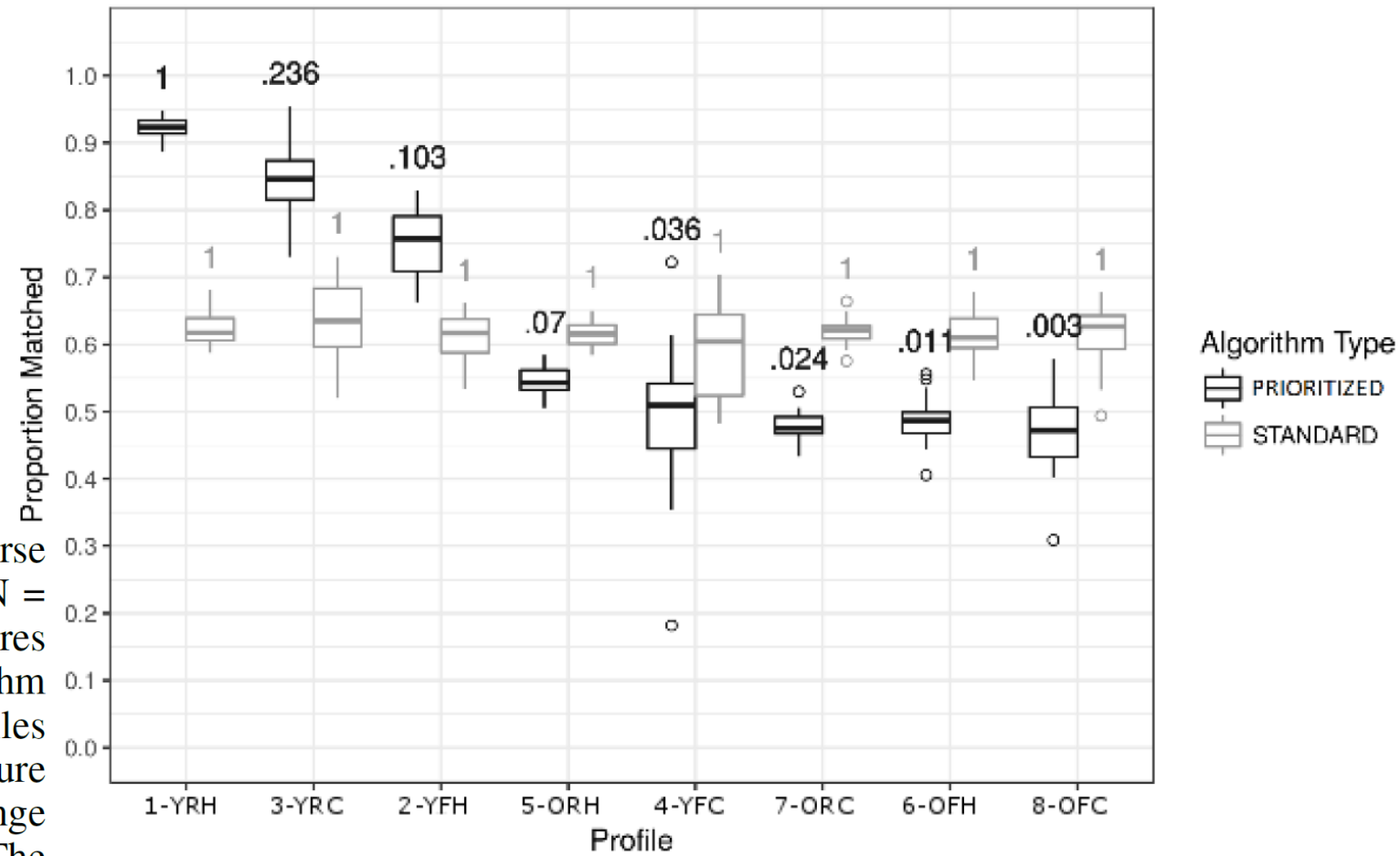
Bradley-Terry model scores

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

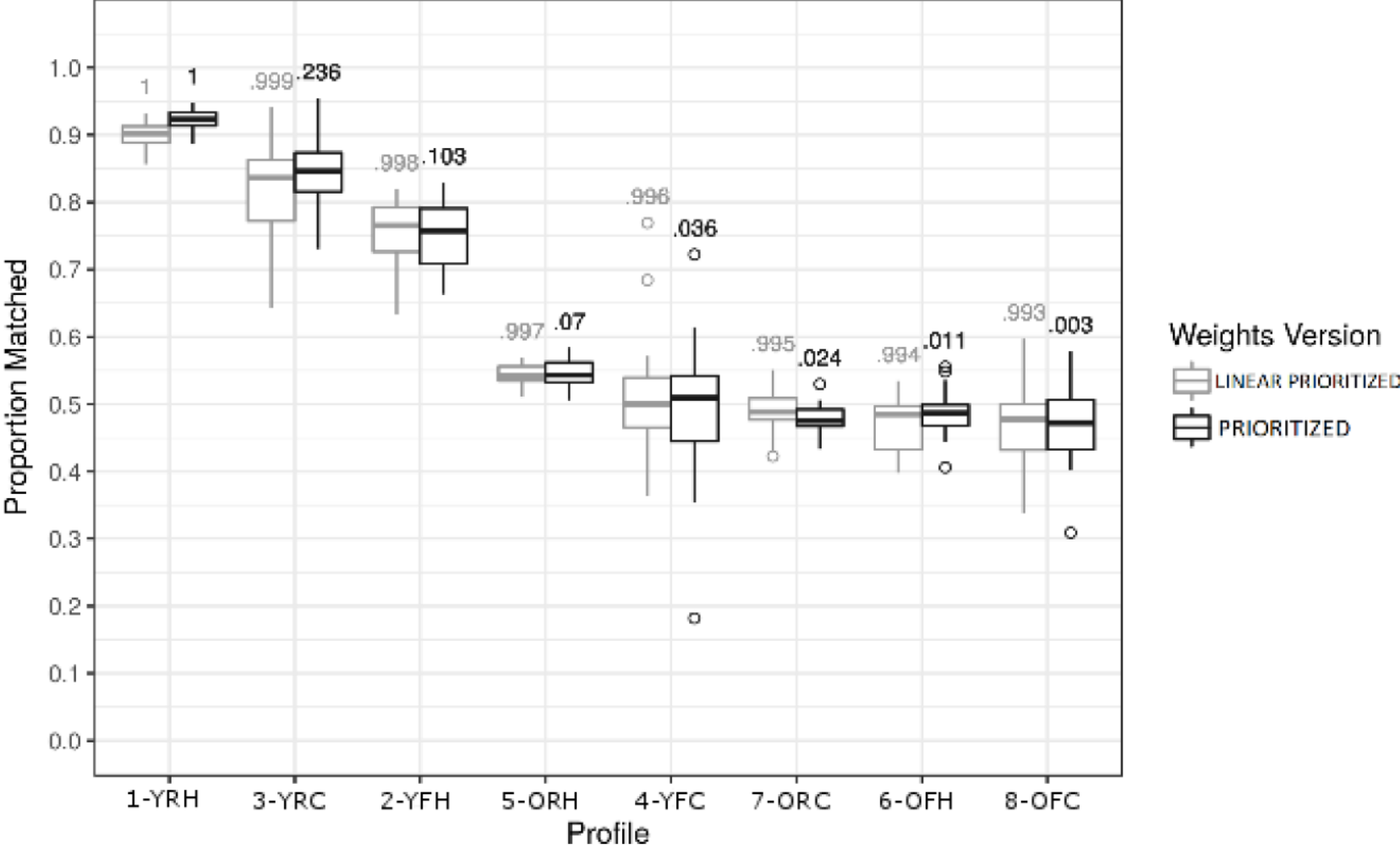
Table 3: The patient profile scores estimated using the Bradley-Terry Model. The “Direct” scores correspond to allowing a separate parameter for each profile (we use these in our simulations below), and the “Attribute-based” scores are based on the attributes via the linear model.

Effect of tiebreaking by profiles

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



Monotone transformations of the weights seem to make little difference



Classes of pairs of blood types

[Ashlagi and Roth 2014; Toulis and Parkes 2015]

- When generating sufficiently large random markets, patient-donor pairs' situations can be categorized according to their blood types
- *Underdemanded* pairs contain a patient with blood type O, a donor with blood type AB, or both
- *Overdemanded* pairs contain a patient with blood type AB, a donor with blood type O, or both
- *Self-demanded* pairs contain a patient and donor with the same blood type
- *Reciprocally demanded* pairs contain one person with blood type A, and one person with blood type B

Most of the effect is felt by underdemanded pairs

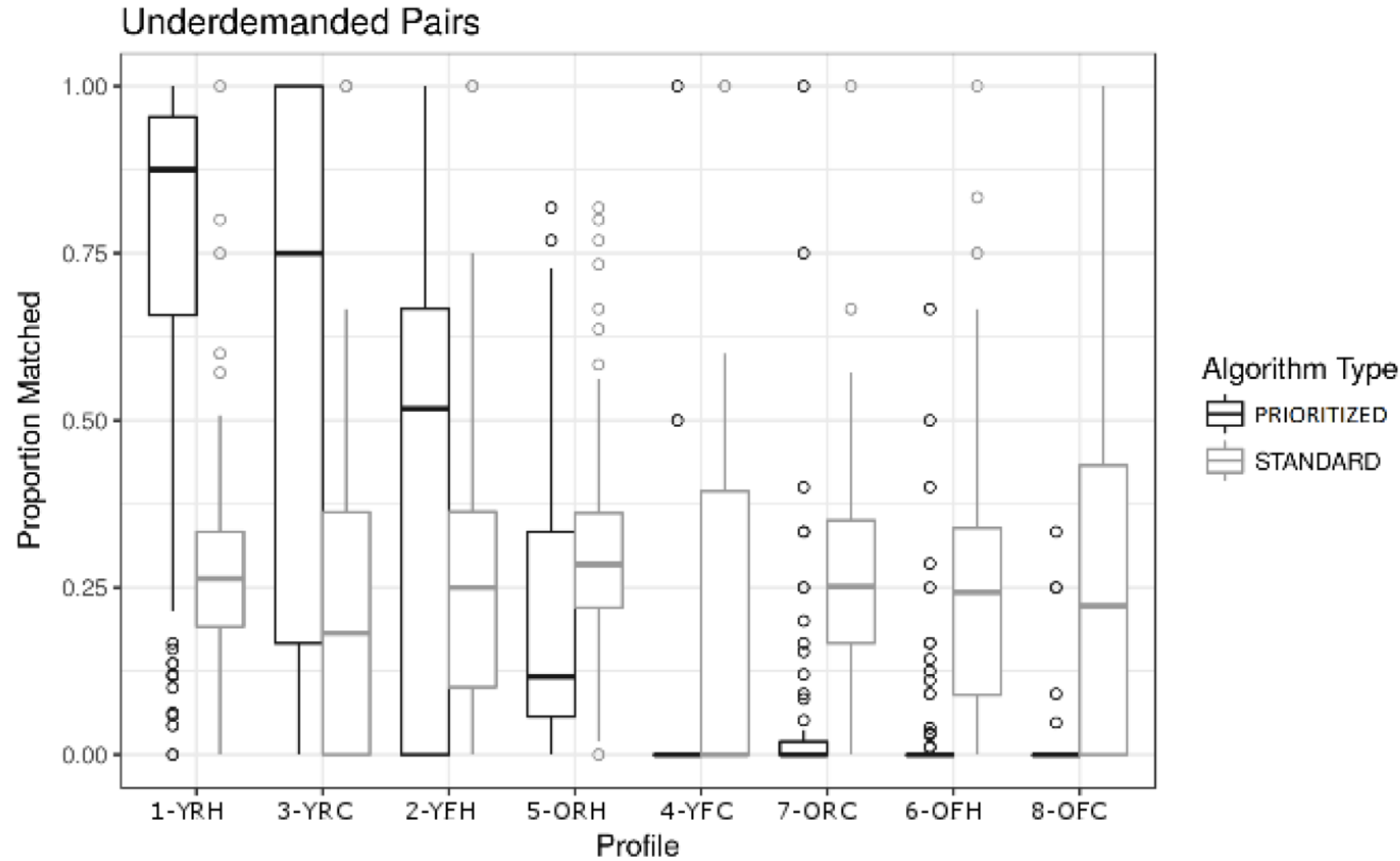


Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.



AAAI/ACM Conference on

**Artificial Intelligence,
Ethics, and Society**

Honolulu, Hawaii, USA

January 27-28, 2019

CALL FOR PAPERS