

Ethical and Societal Worries about AI



autonomous weapons



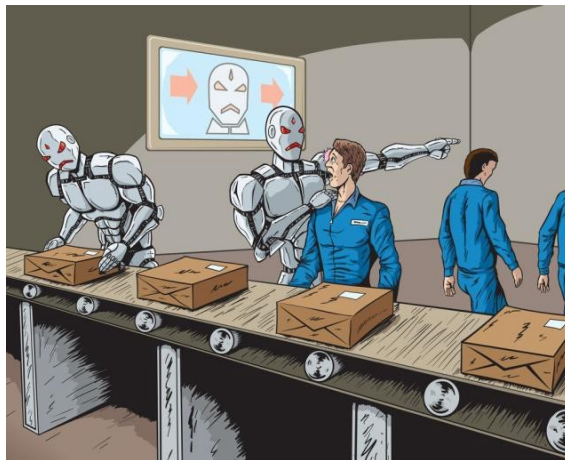
AI & cybersecurity, privacy



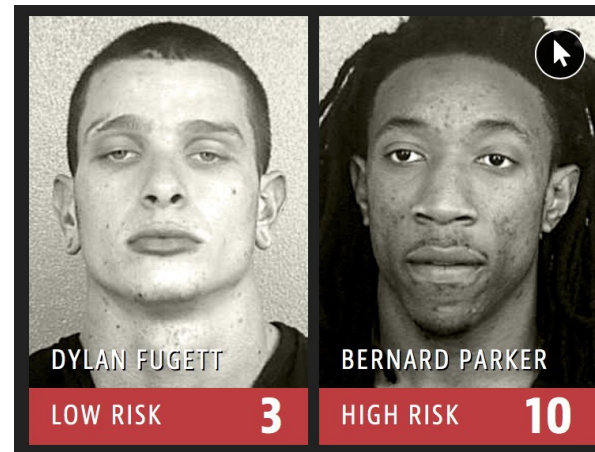
societal surveillance



media manipulation,
polarization



technological unemployment



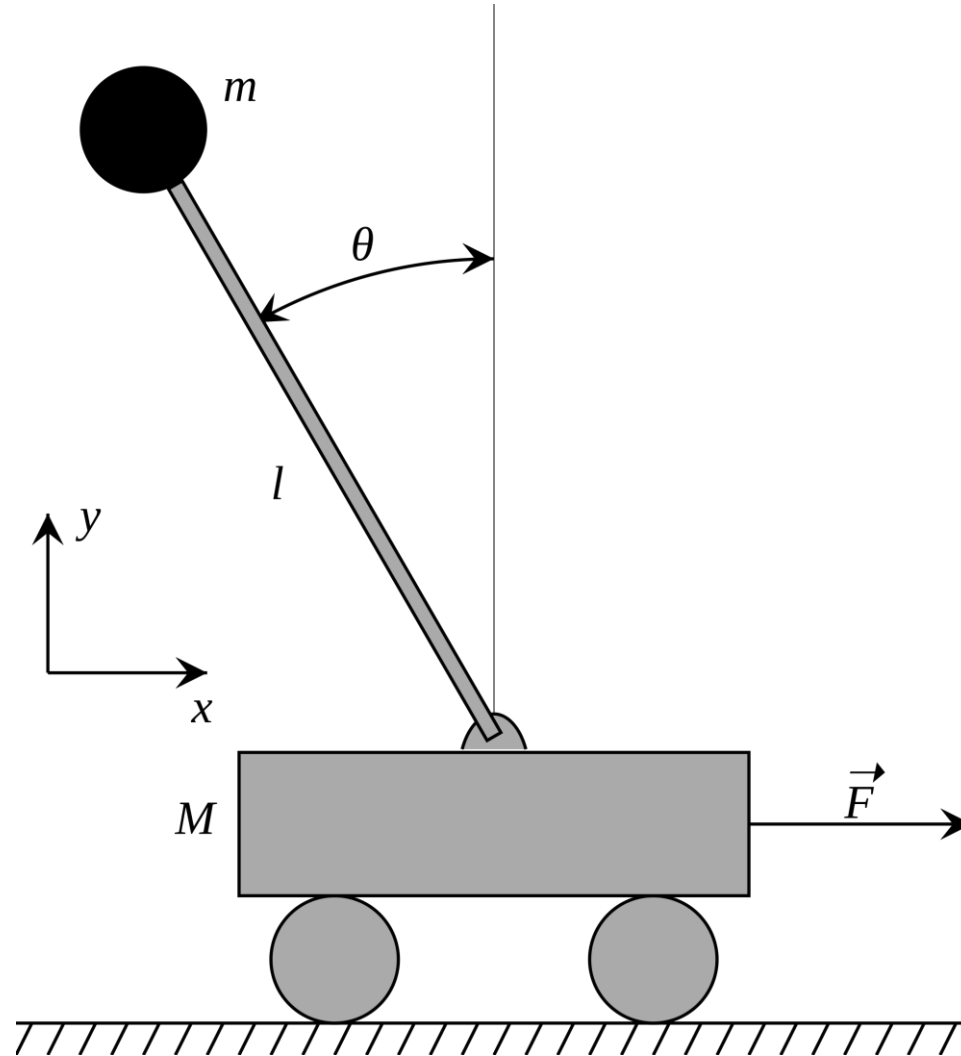
unfair biases



responsibility and liability

...

In the lab, simple objectives are good...



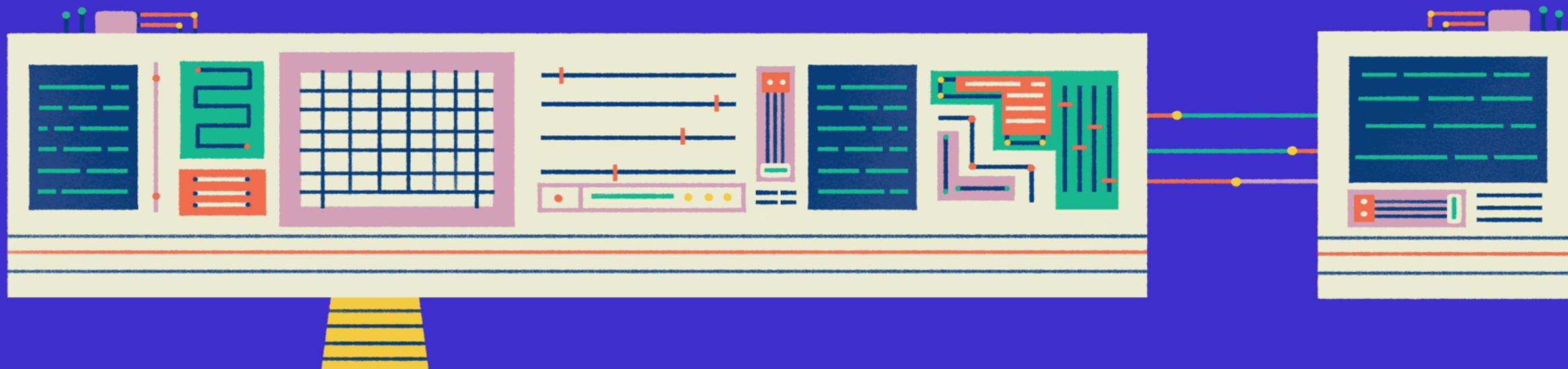
Prescription AI

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

THE BOTPERATING TABLE

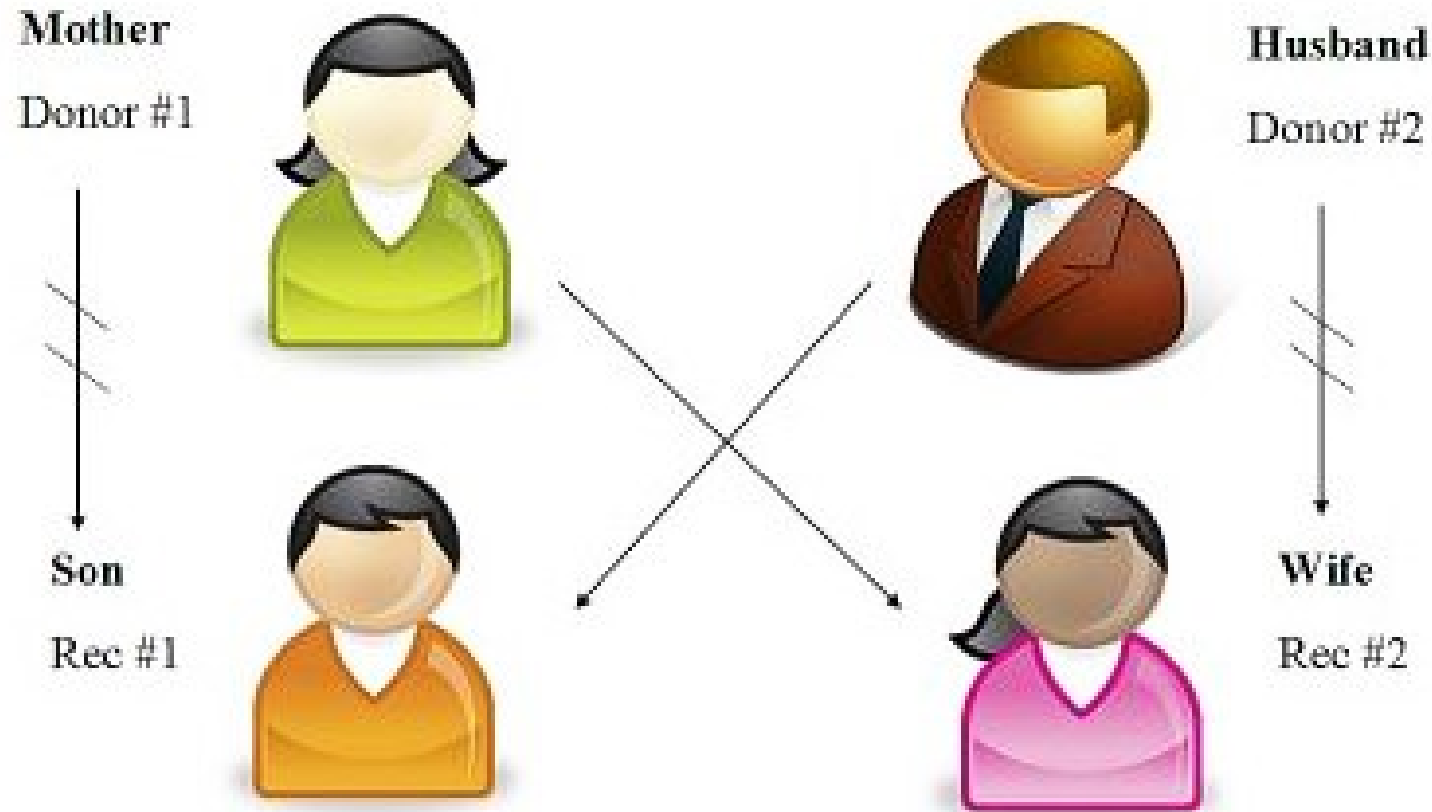
How AI changed organ donation in the US

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

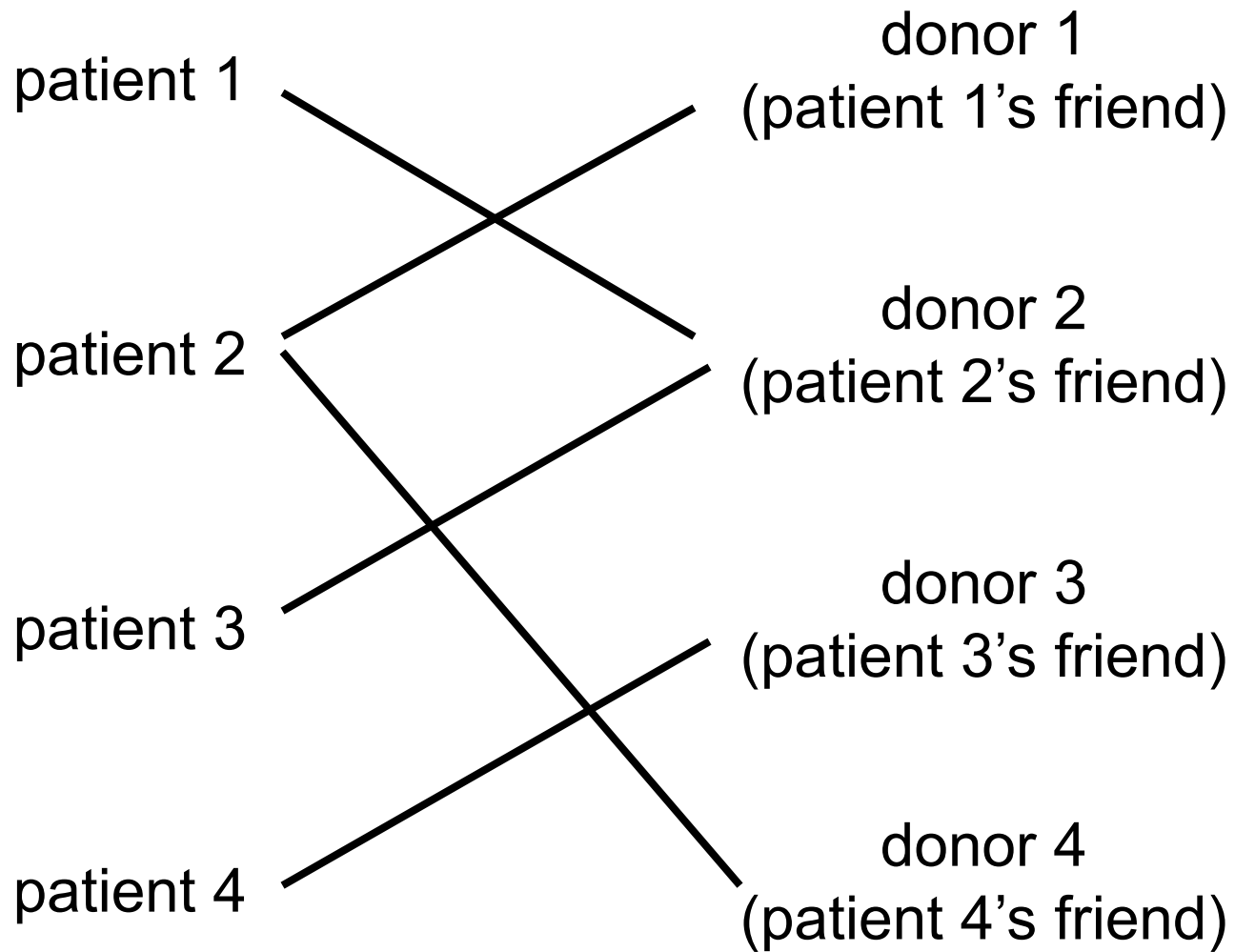


Kidney exchange [Roth, Sönmez, and Ünver 2004]

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



More complex example



Poll 1: which combinations of transplants could we reasonably perform?

A: {p1d2}

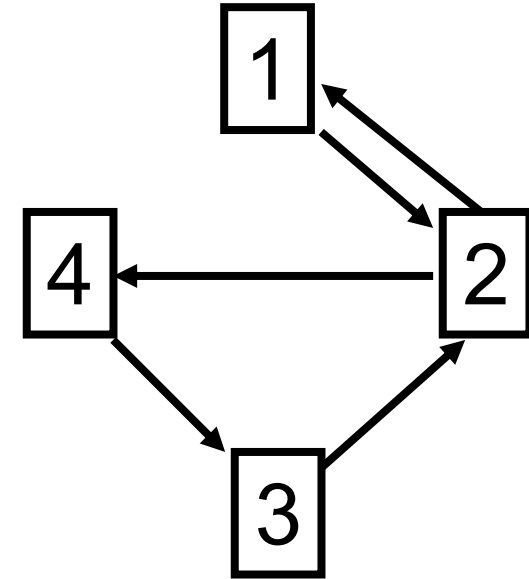
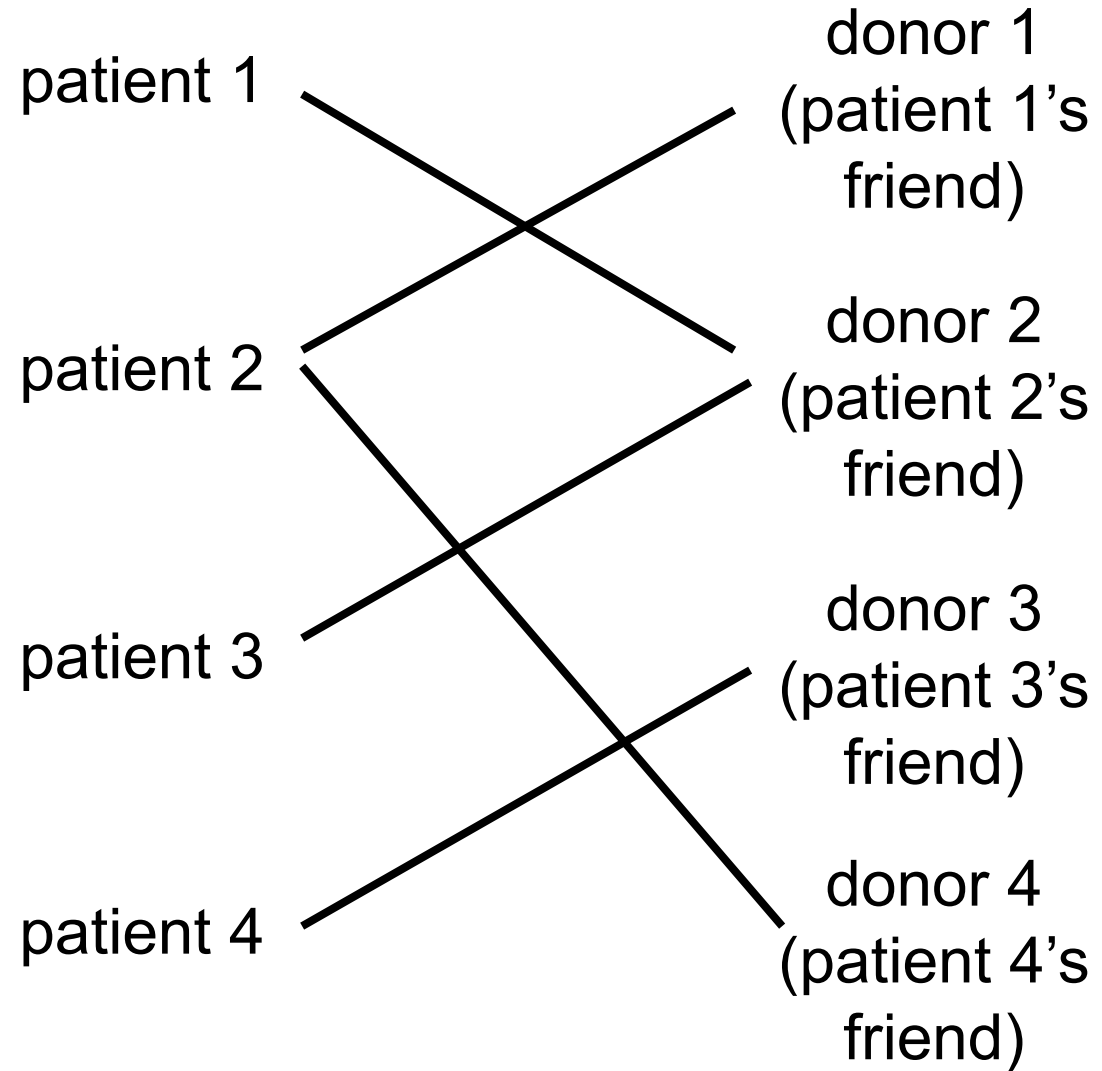
B: {p1d2,p2d1}

C: {p2d4,p3d2}

D: {p2d4,p3d2,p4d3}

E: {p1d2,p2d1,p2d4,p3d2,p4d3}

Different representation



edge from i to j =
patient i wants
donor j 's kidney

Integer programming formulation [\[Abraham, Blum, Sandholm 2007\]](#)

- For each cycle c of length at most k , make a binary variable x_c
 - value 1 if all edges on this cycle are used, 0 otherwise
- maximize $\sum_c |c| x_c$
- subject to:
- for every vertex i : $\sum_{c: i \in c} x_c \leq 1$
 - (every vertex in at most one used cycle)

Adapting a Kidney Exchange Algorithm to Align with Human Values

[AIJ 2020]

with:



Rachel
Freedman



Jana Schaich
Borg



Walter Sinnott-
Armstrong



John P.
Dickerson

Tiebreaking (or more than just tiebreaking?)

- How should we break ties?
 - (Should we do more than break ties?)
 - Who should decide? How? What information would they need?

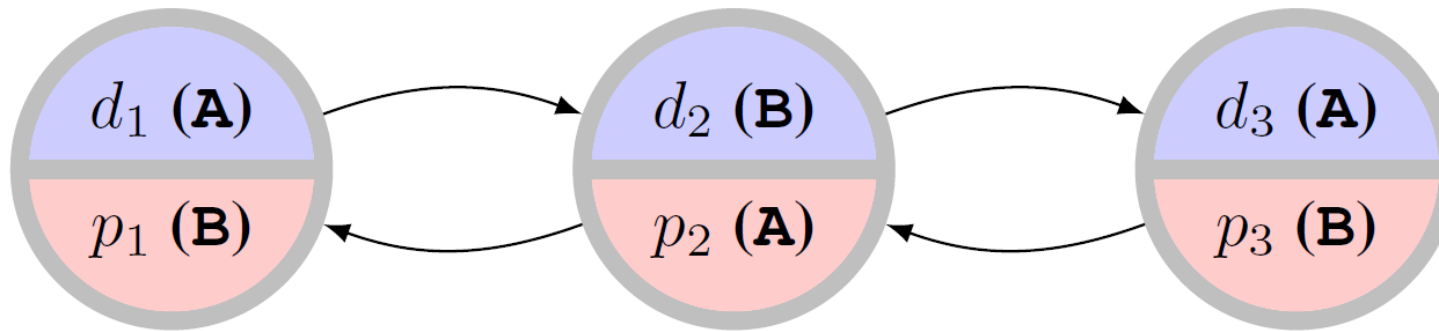


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

Eliciting attributes

Table 2

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

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

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.

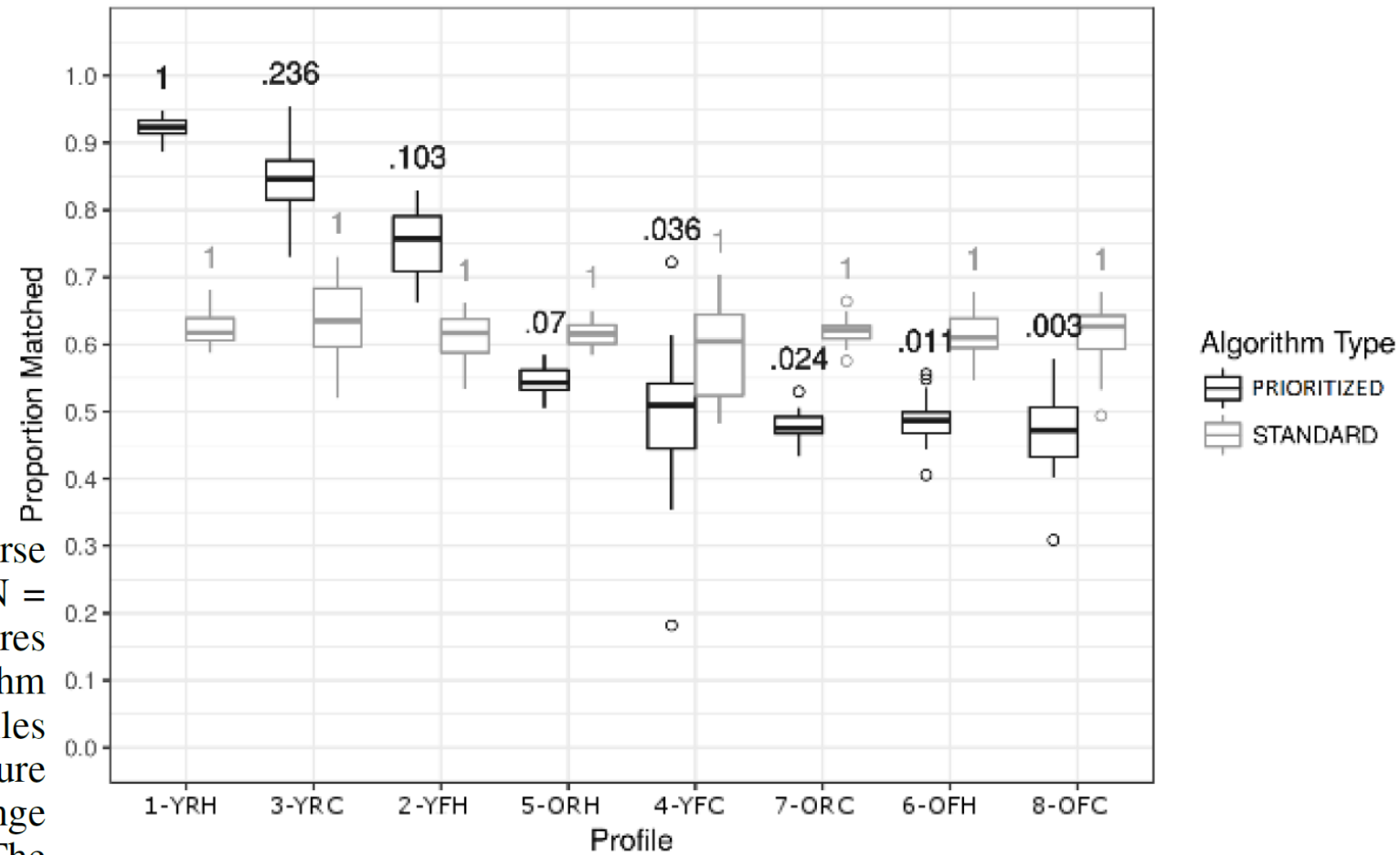
Bradley-Terry model scores

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.



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

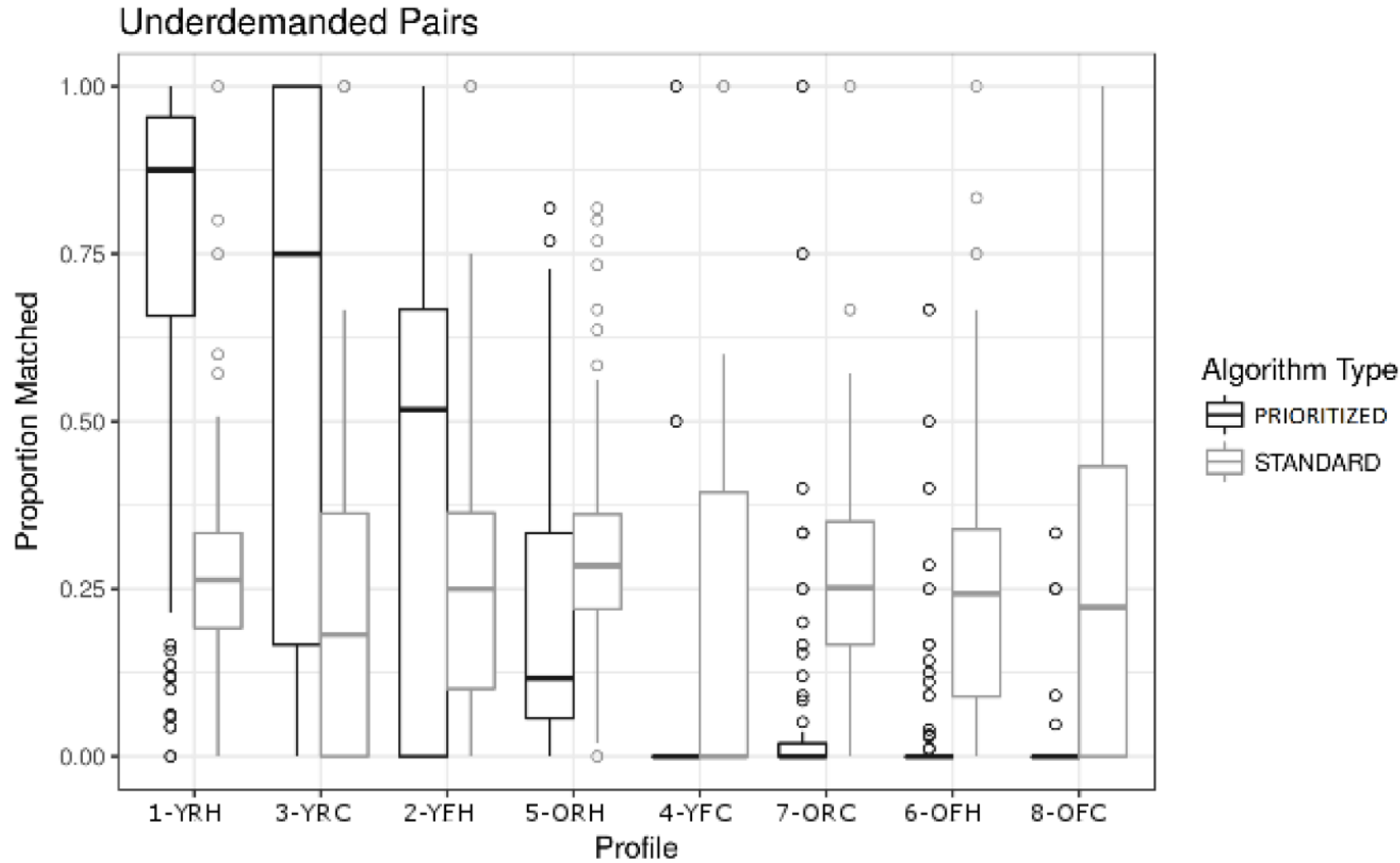
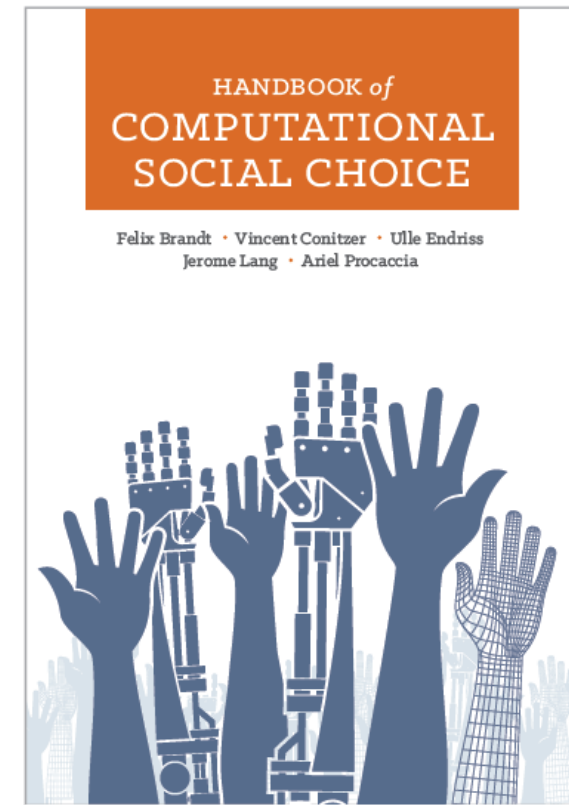


Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

Concerns about learning from people

- What if we predict people will disagree?
 - New social-choice theoretic questions [C. et al. 2017] – approach also followed by Noothigattu et al. [2018], Kahng et al. [2019]
- This will *at best* result in current human-level moral decision making [raised by, e.g., Chaudhuri and Vardi 2014]
 - ... though might perform better than any *individual* person because individual's errors are voted out
- How to generalize appropriately? Representation?



Fifth AAAI /ACM Conference on
**Artificial Intelligence,
Ethics, and Society**
Oxford
August 1-3, 2022



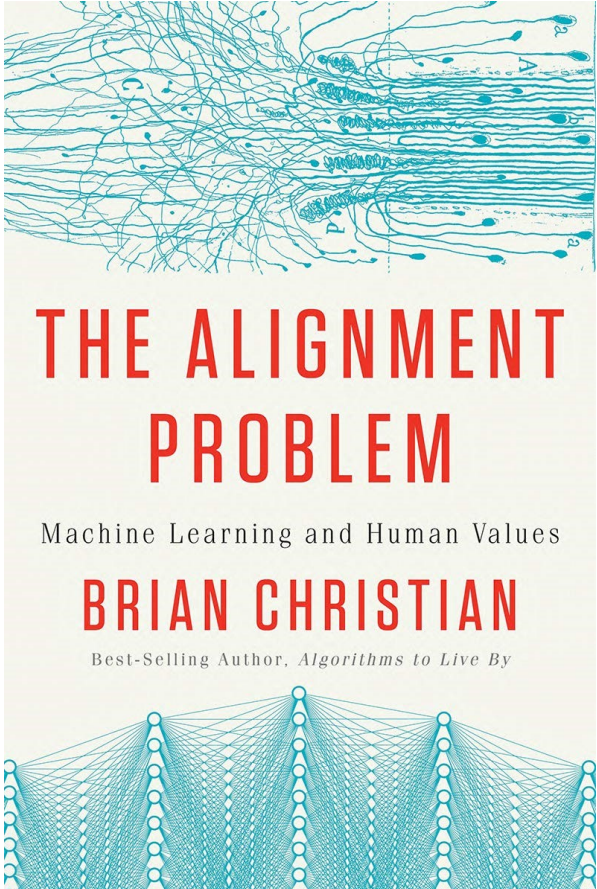
FAccT
서울
2022



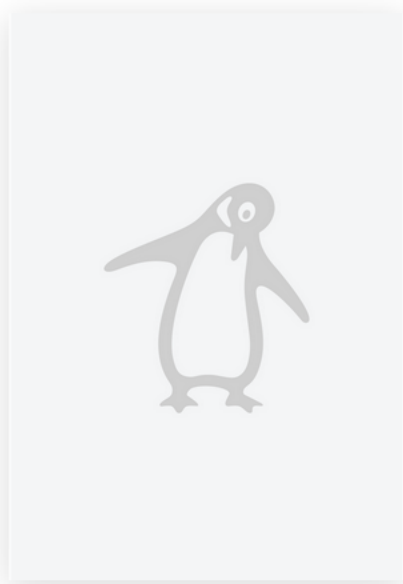
**Institute for
Ethics in AI**
Oxford leading the way in AI
ethics

Stanford University

One Hundred Year Study on Artificial
Intelligence (AI100)



**THE ALIGNMENT
PROBLEM**
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-Selling Author, *Algorithms to Live By*



Released 08/02/2024

Details +

All Editions +

Share +

Jana Schaich Borg, Walter Sinnott-Armstrong, Vincent Conitzer

Moral AI

And How We Get There

HARDBACK



PRE-ORDER



Summary

A reassuring and thought-provoking guide to all the big questions about AI and ethics

Should robots ever be considered free? Will computers transcend human intelligence? And what can we do to make sure AI is safe?

The artificial intelligence revolution has begun. Today, there are self-driving cars on our streets, autonomous weapons in our armies, robot surgeons in our hospitals - and AI's presence in our lives will only increase. Some see this as the dawn of new era in innovation and ease; others are alarmed by its destructive potential. But one thing is clear: this is a technology like no other, one that raises profound questions about freedom, justice and the very definition of human agency.