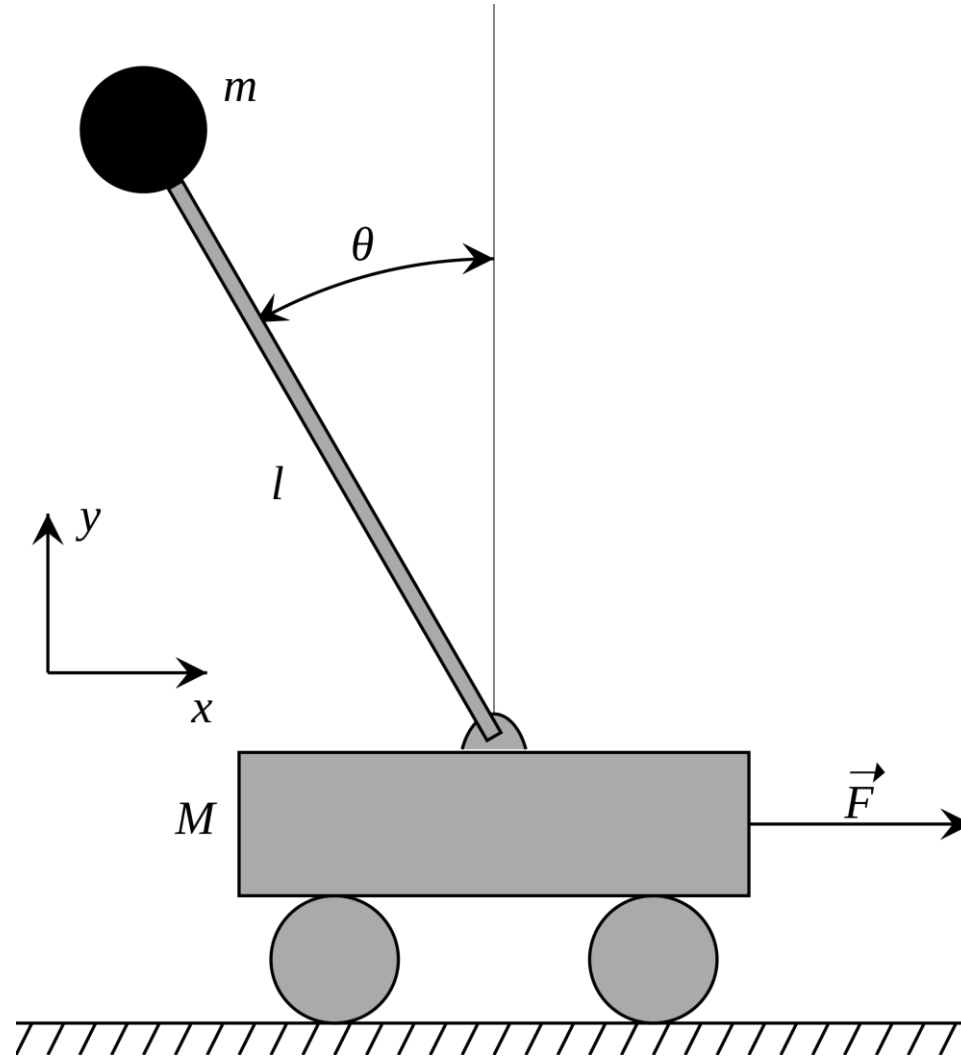


Automated Moral Decision Making

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.

...

Ethical and Societal Worries about AI



autonomous weapons



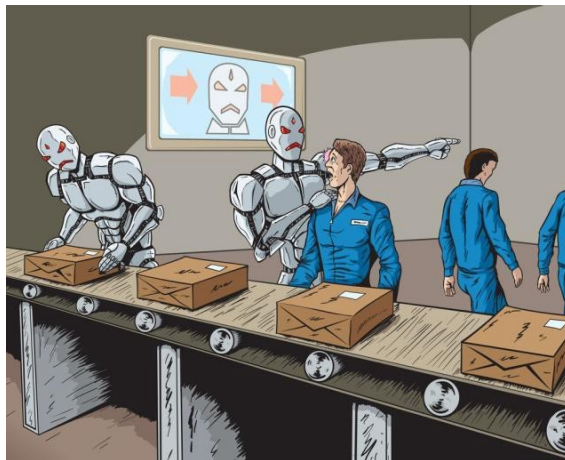
AI & cybersecurity, privacy



societal surveillance



media manipulation,
polarization



technological unemployment



unfair biases



responsibility and liability

...



Fifth AAAI /ACM Conference on

Artificial Intelligence, Ethics, and Society

Oxford

August 1-3, 2022

Moral Decision Making Frameworks for Artificial Intelligence

[AAAI'17]

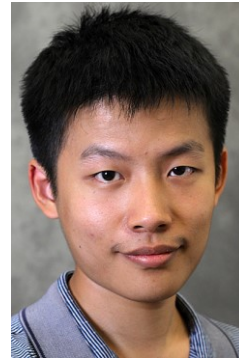
with:



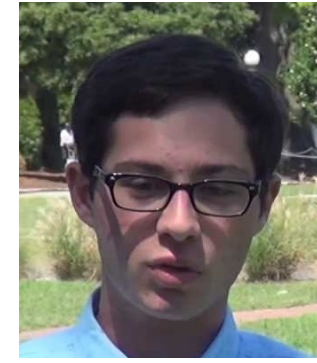
Walter Sinnott-
Armstrong



Jana Schaich
Borg



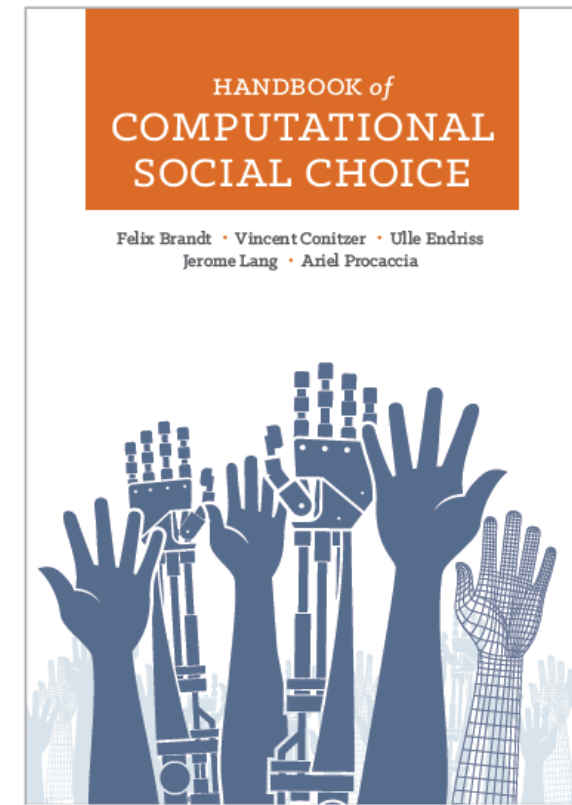
Yuan Deng



Max Kramer

Concerns with 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?



Social-choice-theoretic approaches

- C., Sinnott-Armstrong, Schaich Borg, Deng, Kramer [AAAI'17]: “[give] the AI some type of social-choice-theoretic aggregate of the moral values that we have inferred (for example, by letting our models of multiple people’s moral values *vote* over the relevant alternatives, or using only the moral values that are common to all of them).”
- C., Schaich Borg, Sinnott-Armstrong [Trustworthy Algorithmic Decision Making Workshop'17]: “One possible solution is to let the models of multiple subjects *vote* over the possible choices. But exactly how should this be done? Whose preferences should count and what should be the voting rule used? How do we remove bias, prejudice, and confusion from the subjects’ judgments? These are novel problems in computational social choice.”
- Noothigattu, Gaikwad, Awad, Dsouza, Rahwan, Ravikumar, Procaccia [AAAI'18]:
 - **I. Data collection:** Ask human voters to compare pairs of alternatives (say a few dozen per voter). In the autonomous vehicle domain, an alternative is determined by a vector of features such as the number of victims and their gender, age, health — even species!
 - **II. Learning:** Use the pairwise comparisons to learn a model of the preferences of each voter over all possible alternatives.
 - **III. Summarization:** Combine the individual models into a single model, which approximately captures the collective preferences of all voters over all possible alternatives.
 - **IV. Aggregation:** At runtime, when encountering an ethical dilemma involving a specific subset of alternatives, use the summary model to deduce the preferences of all voters over this particular subset, and apply a voting rule to aggregate these preferences into a collective decision.”
- Kahng, Lee, Noothigattu, Procaccia, Psomas [ICML'19]: The idea is that we would ideally like to consult the voters on each decision, but in order to automate those decisions we instead use the models that we have learned as a proxy for the flesh and blood voters. In other words, the models serve as virtual voters, which is why we refer to this paradigm as *virtual democracy*.

Scenarios

- You see a woman throwing a stapler at her colleague who is snoring during her talk. How morally wrong is the action depicted in this scenario?
 - Not at all wrong (1)
 - Slightly wrong (2)
 - Somewhat wrong (3)
 - Very wrong (4)
 - Extremely wrong (5)

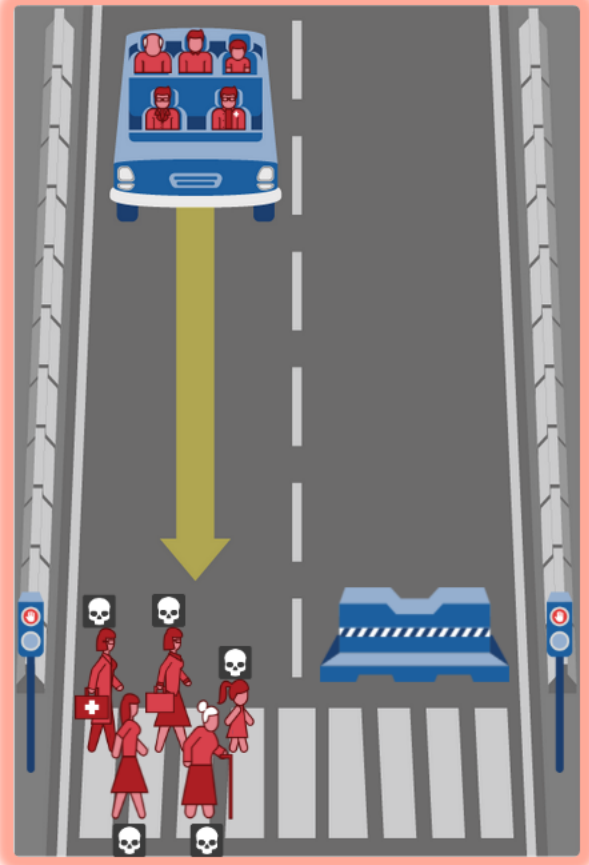
[Clifford, Iyengar, Cabeza, and Sinnott-Armstrong, "Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory." *Behavior Research Methods*, 2015.]

What should the self-driving car do?

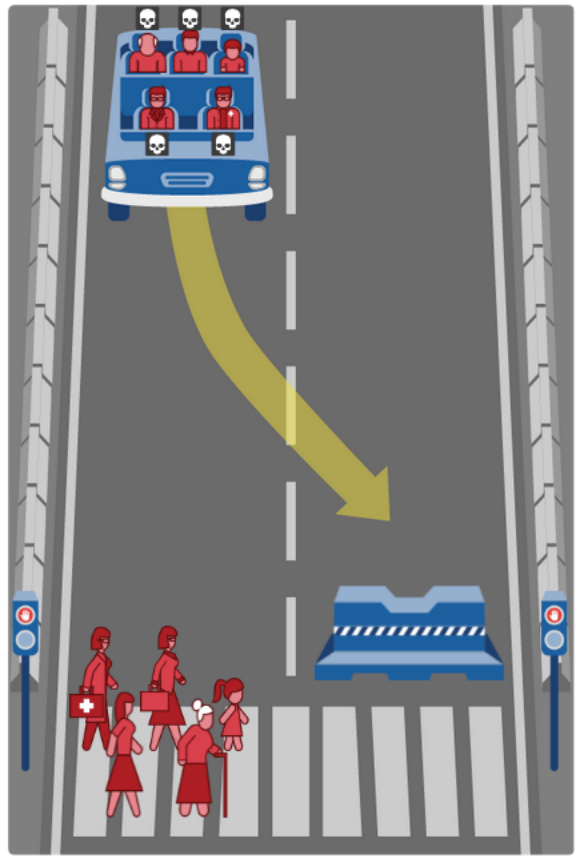
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- 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.

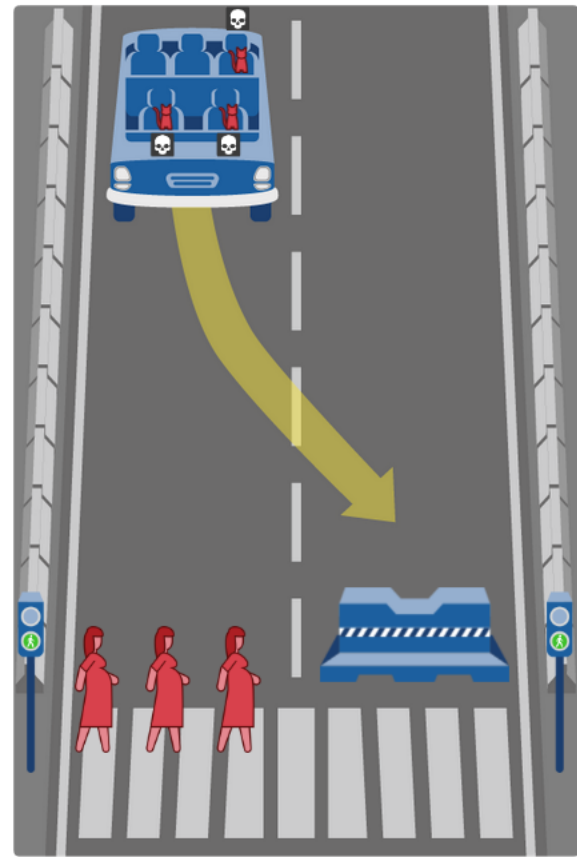
Bonnefon, Shariff, Rahwan, "The social dilemma of autonomous vehicles." *Science* 2016

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.



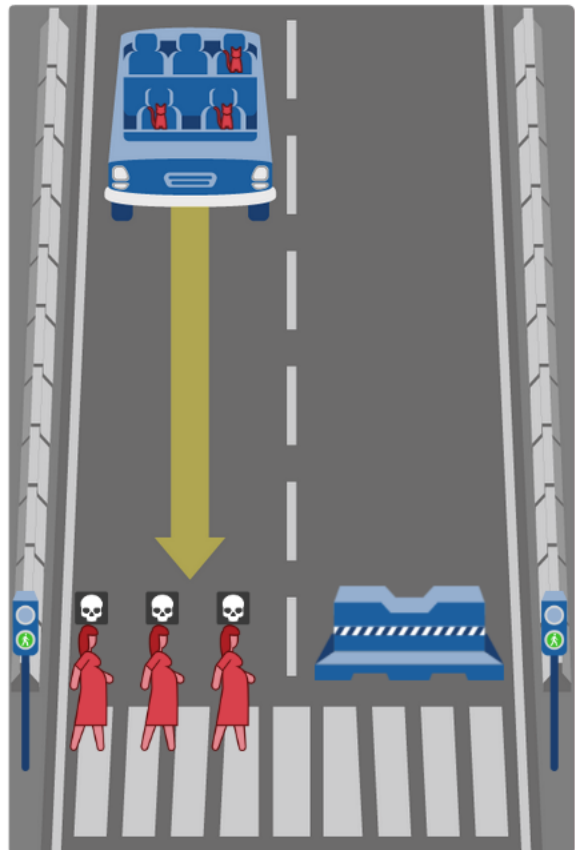
Hide Description

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

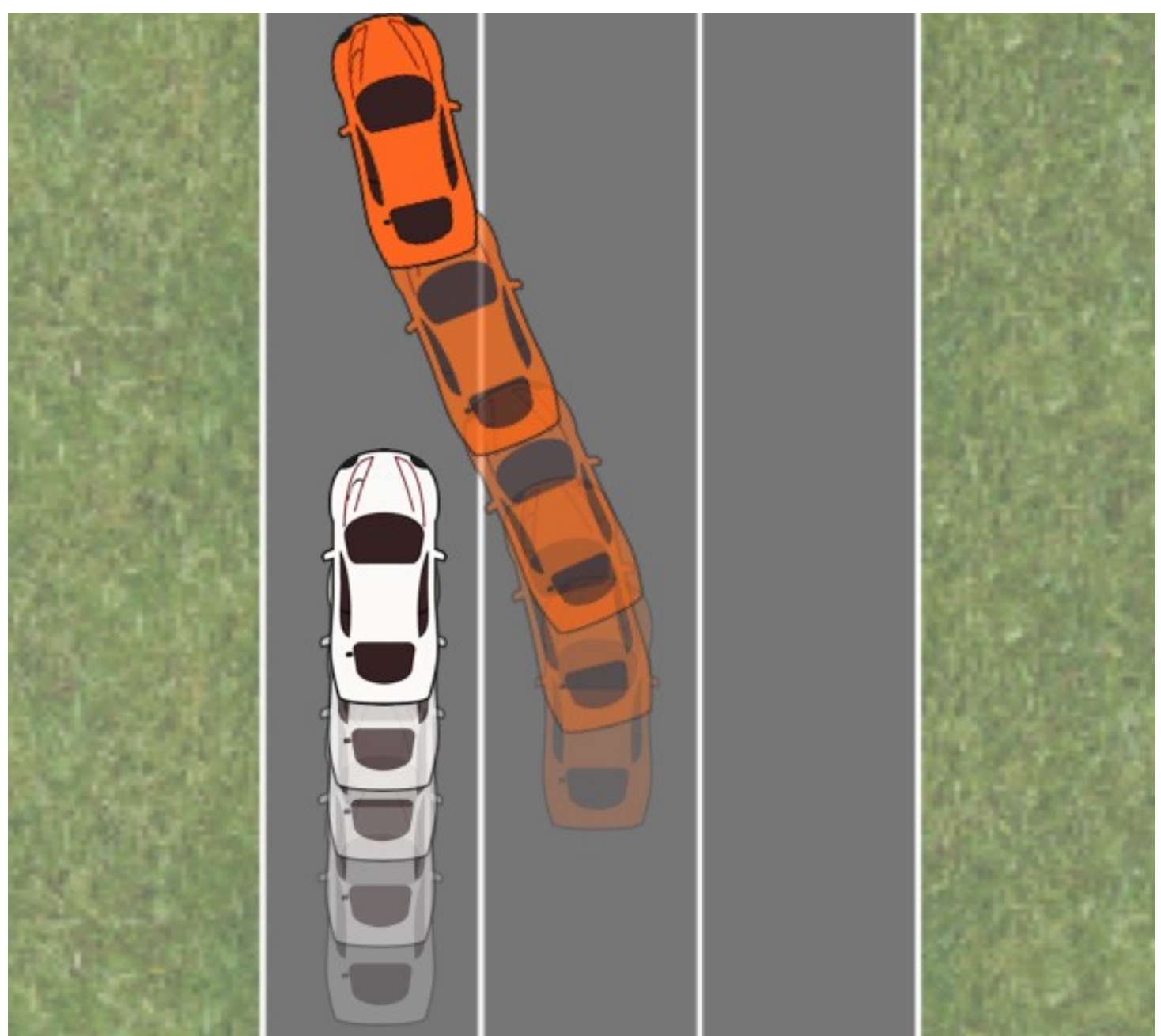


Protecting Passengers



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

[Artificial Intelligence (AIJ) 2020]

with:



Rachel
Freedman



Jana Schaich
Borg



Walter Sinnott-
Armstrong



John P.
Dickerson

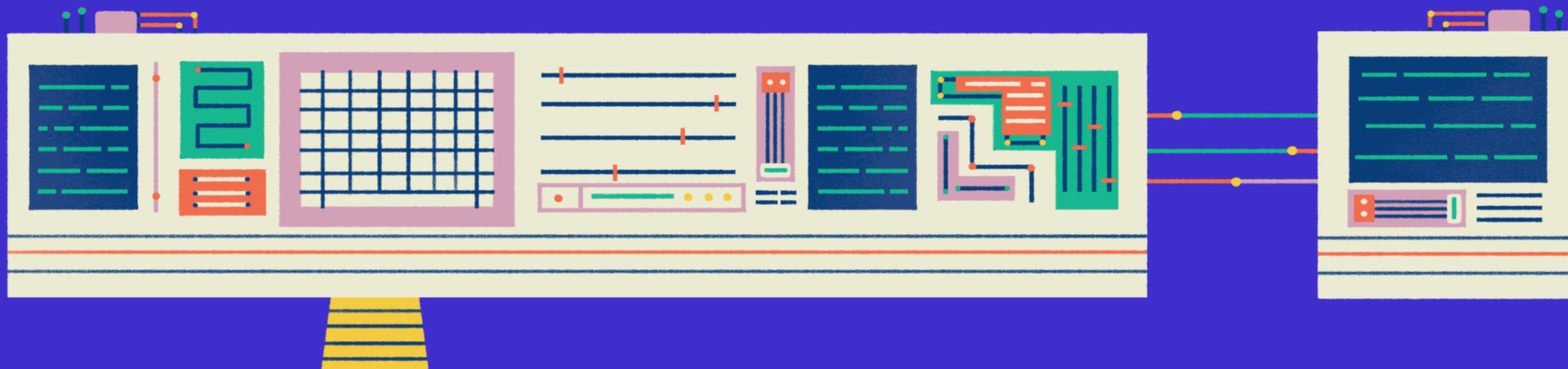
Prescription AI

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

THE BOTPERATING TABLE

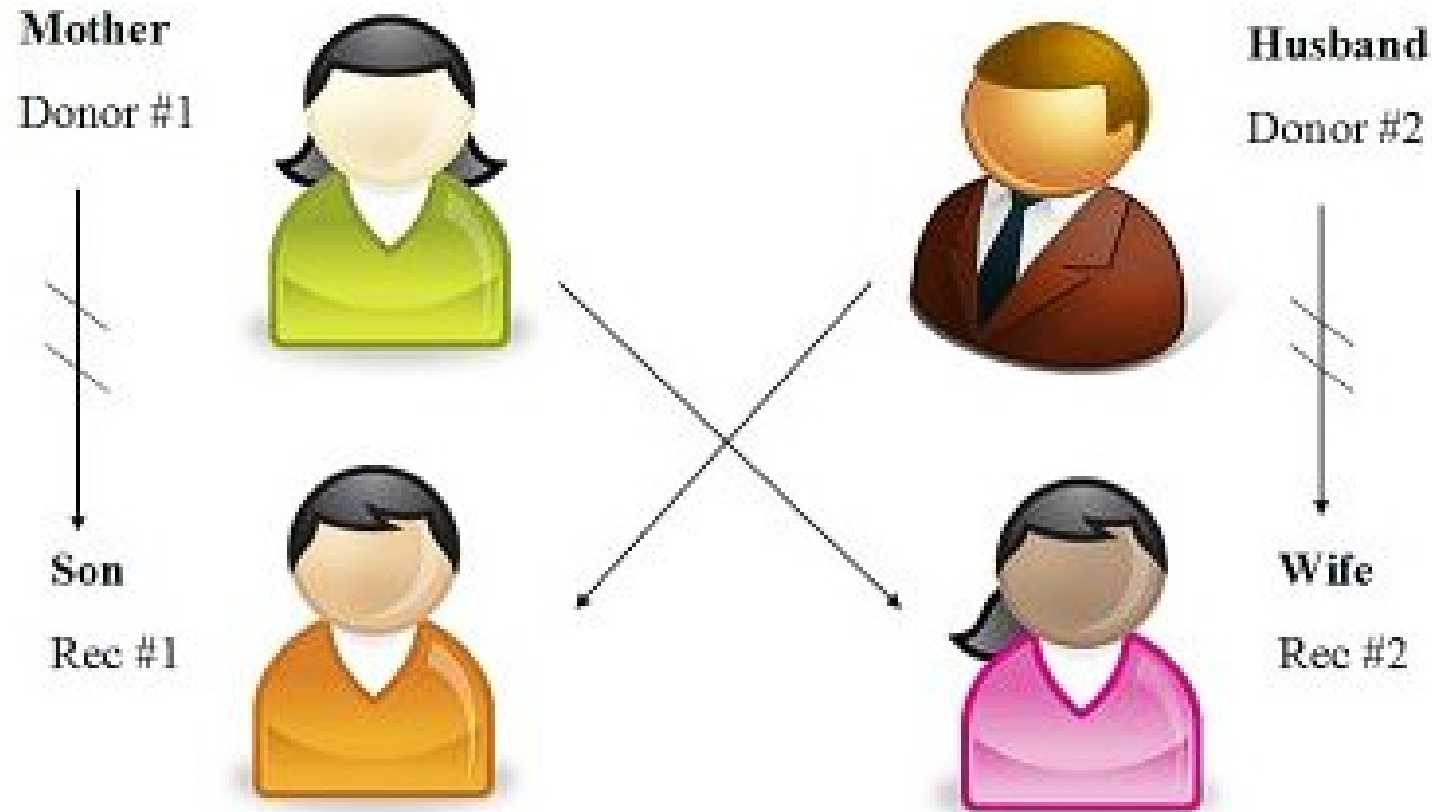
How AI changed organ donation in the US

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

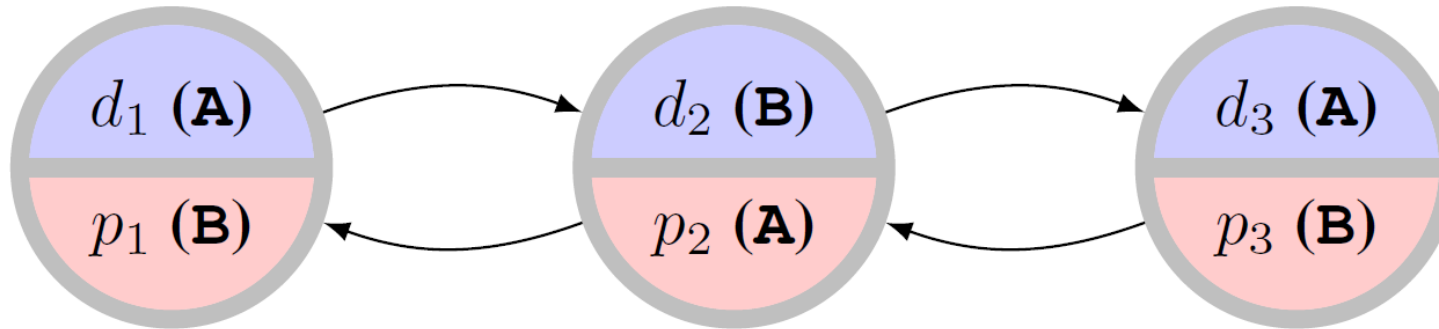


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

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

Another example

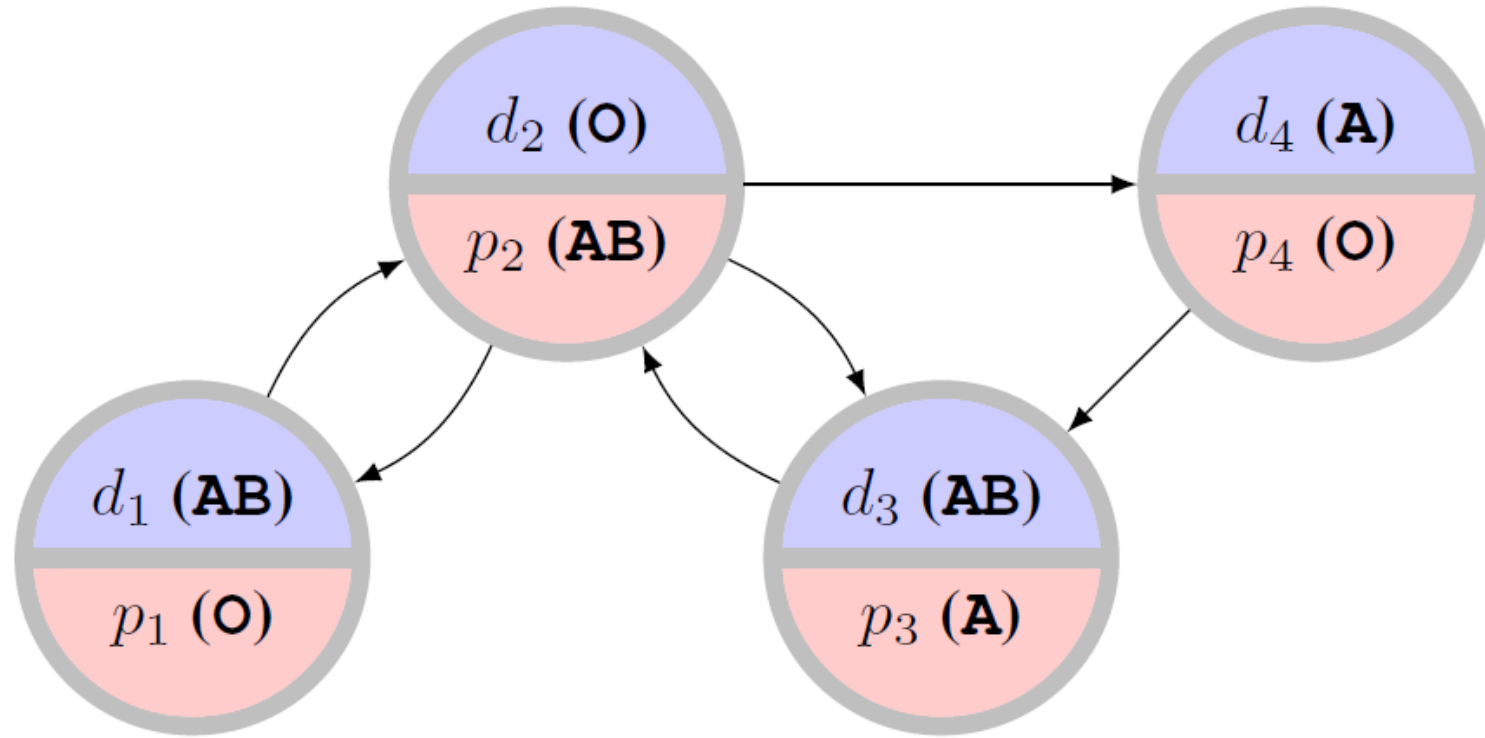


Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

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.

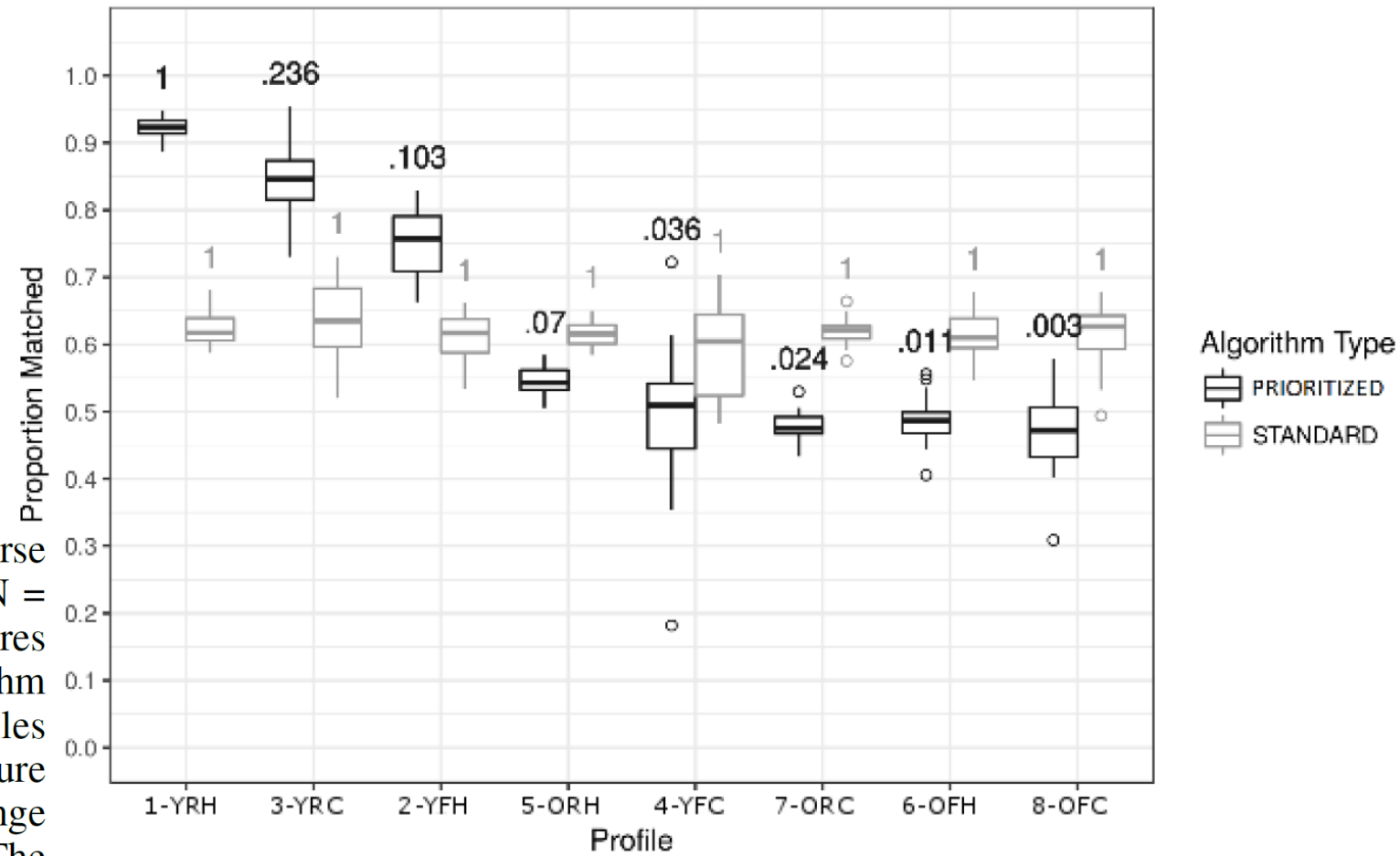
Bradley-Terry model scores

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

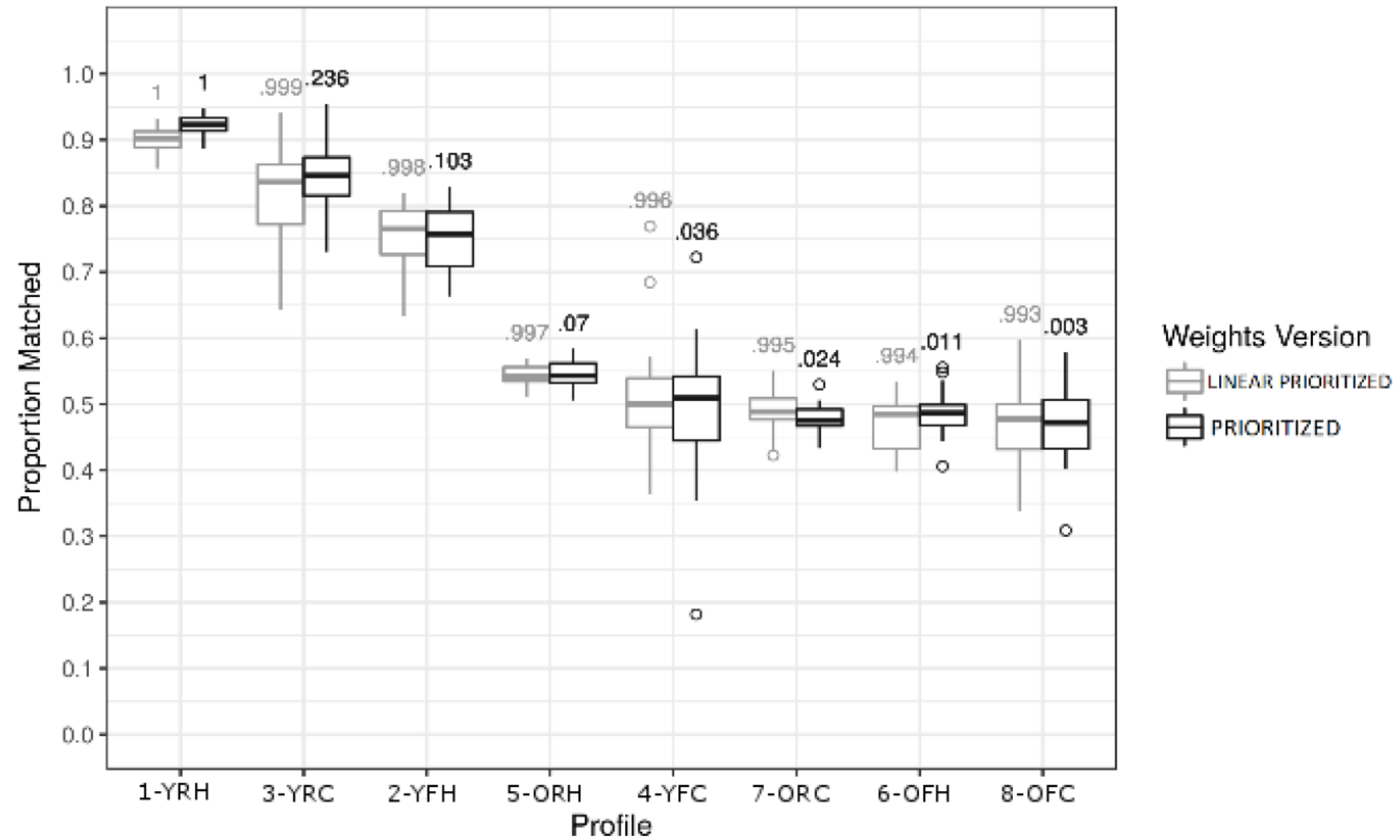
Table 3: The patient profile scores estimated using the Bradley-Terry Model. The “Direct” scores correspond to allowing a separate parameter for each profile (we use these in our simulations below), and the “Attribute-based” scores are based on the attributes via the linear model.

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.



Monotone transformations of the weights 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

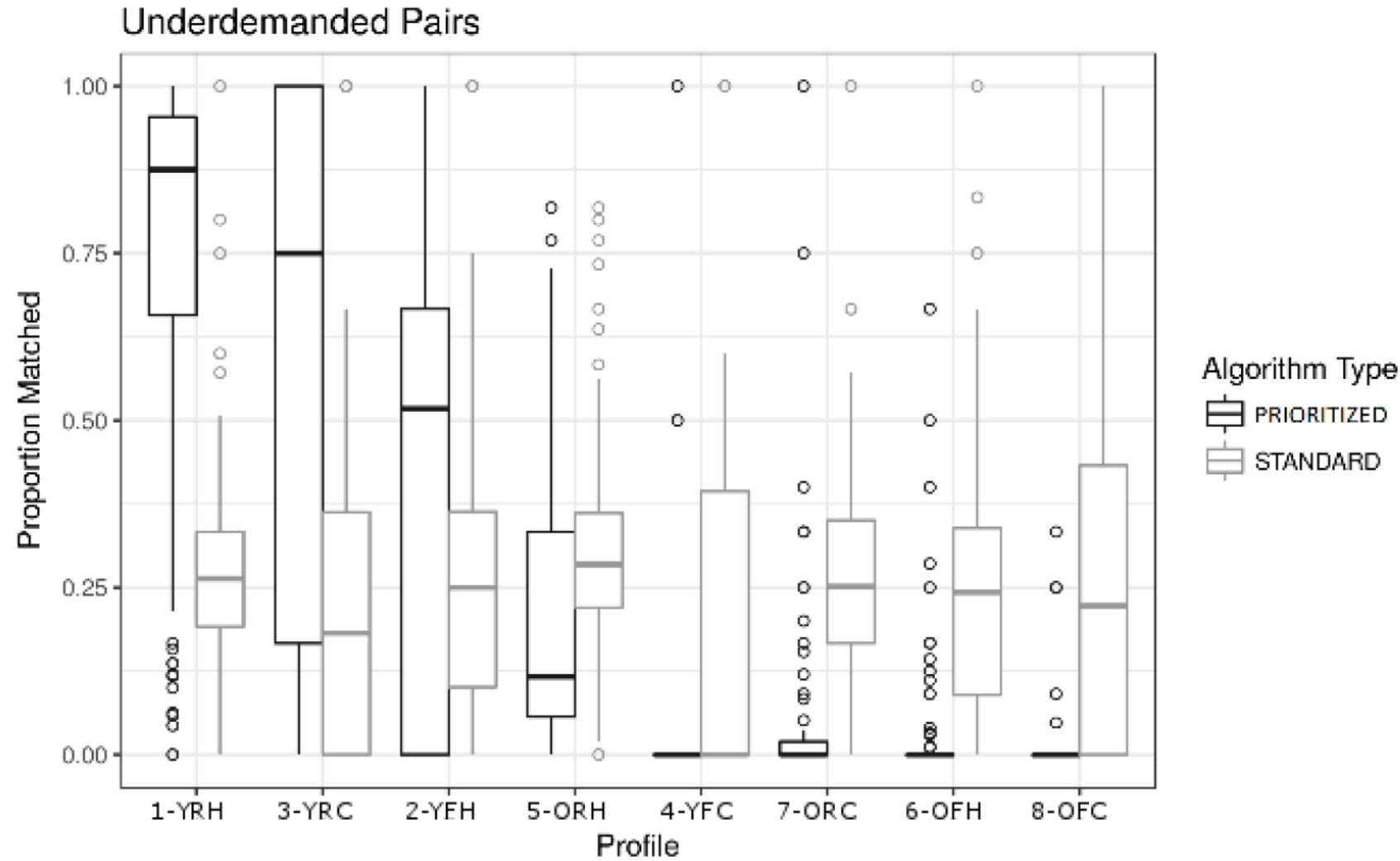
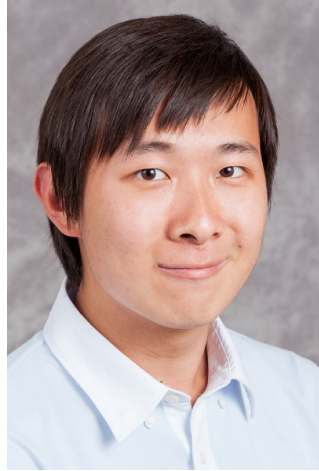


Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

A PAC Learning Framework for Aggregating Agents' Judgments [AAAI'19]

with:



Hanrui
Zhang

How many subjects do we
need to query?

How many queries do we
need to ask each of them?

Learning from agents' judgments

features (e.g., is the patient on the left younger?)

label (e.g., should we prefer the patient on the left?)

Agent	x_1	x_2	x_3	y
Alice	1	0	0	1
Alice	1	0	1	1
Alice	1	1	0	1
Bob	1	0	0	0
Bob	1	0	1	1
Bob	0	0	1	0
Charlie	1	0	0	0
Charlie	1	1	0	1
Charlie	0	0	1	0

conjunctions that fit individuals perfectly

x_1

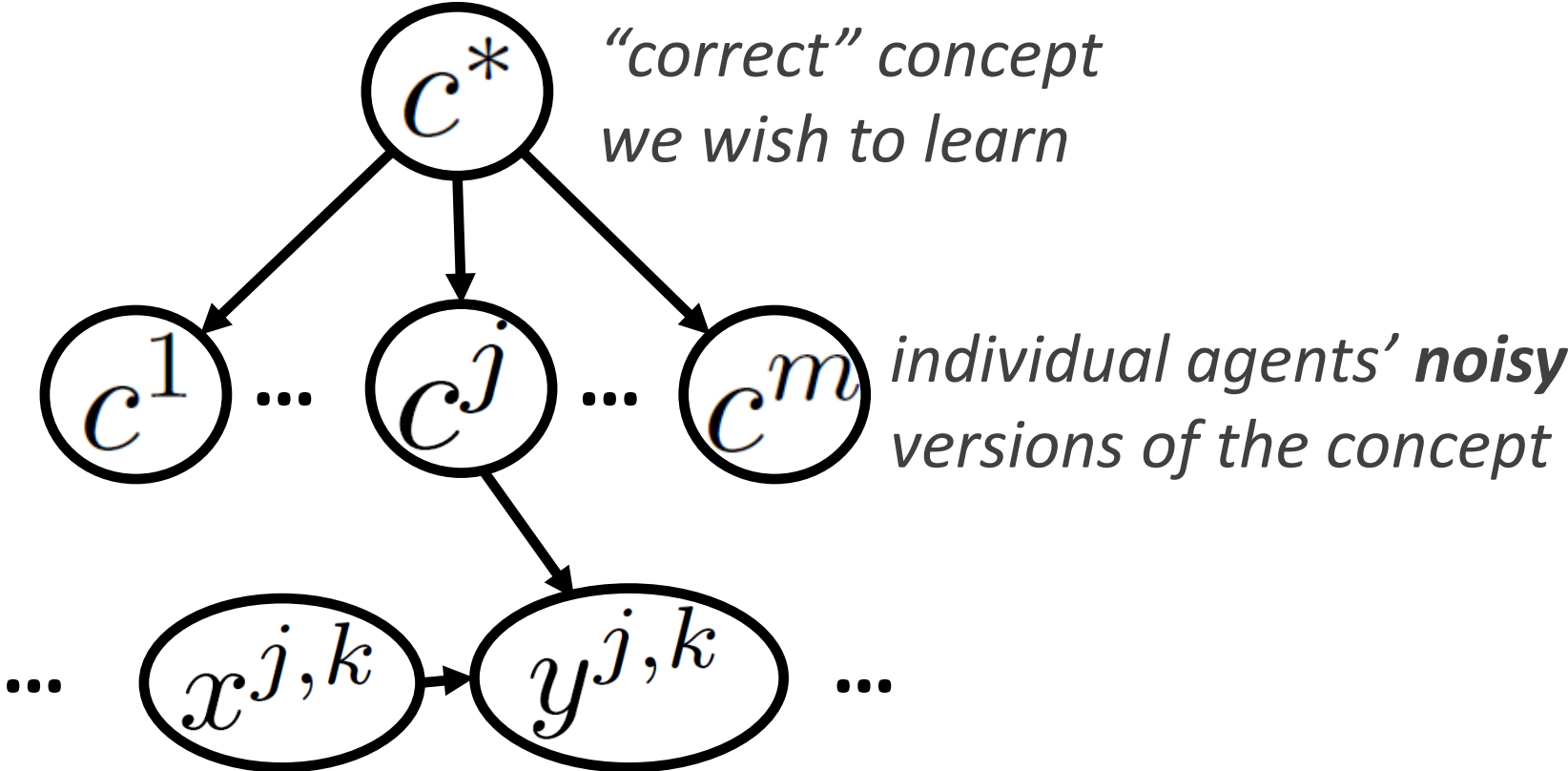
$x_1 \wedge x_3$

x_2

conjunction that fits all data best (two mistakes)

x_1

Our model



*“correct” concept
we wish to learn*

*individual agents’ **noisy**
versions of the concept*

*feature values of
individual example
shown to agent j*

*label given to this
example by j (according
to noisy concept)*

Theorem 3 (Binary Judgments, I.I.D. Symmetric Distributions). *Suppose that $\mathcal{C} = \{-1, 1\}^n$; for each $i \in [n]$, $\mathcal{D}_i = \mathcal{D}_0$ is a non-degenerate⁷ symmetric distribution with bounded absolute third moment; and the noisy mapping with noise rate η satisfies*

$$\nu(c)_i = \begin{cases} c_i, & \text{w.p. } 1 - \eta \\ -1, & \text{w.p. } \eta/2 \\ 1, & \text{w.p. } \eta/2 \end{cases},$$

Then, Algorithm 1 with $m = O\left(\frac{\ln(n/\delta)}{(1-\eta)^2}\right)$ agents and $\ell m = O\left(\frac{n \ln(n/\delta)}{(1-\eta)^2}\right)$ data points in total outputs the correct concept $h = c^$ with probability at least $1 - \delta$.*

Artificial Artificial Intelligence: Measuring Influence of AI "Assessments" on Moral Decision-Making

[AI, Ethics, and Society (AIES) Conference'20]

with:



Lok
Chan



Kenzie
Doyle



Duncan
McElfresh



John P.
Dickerson

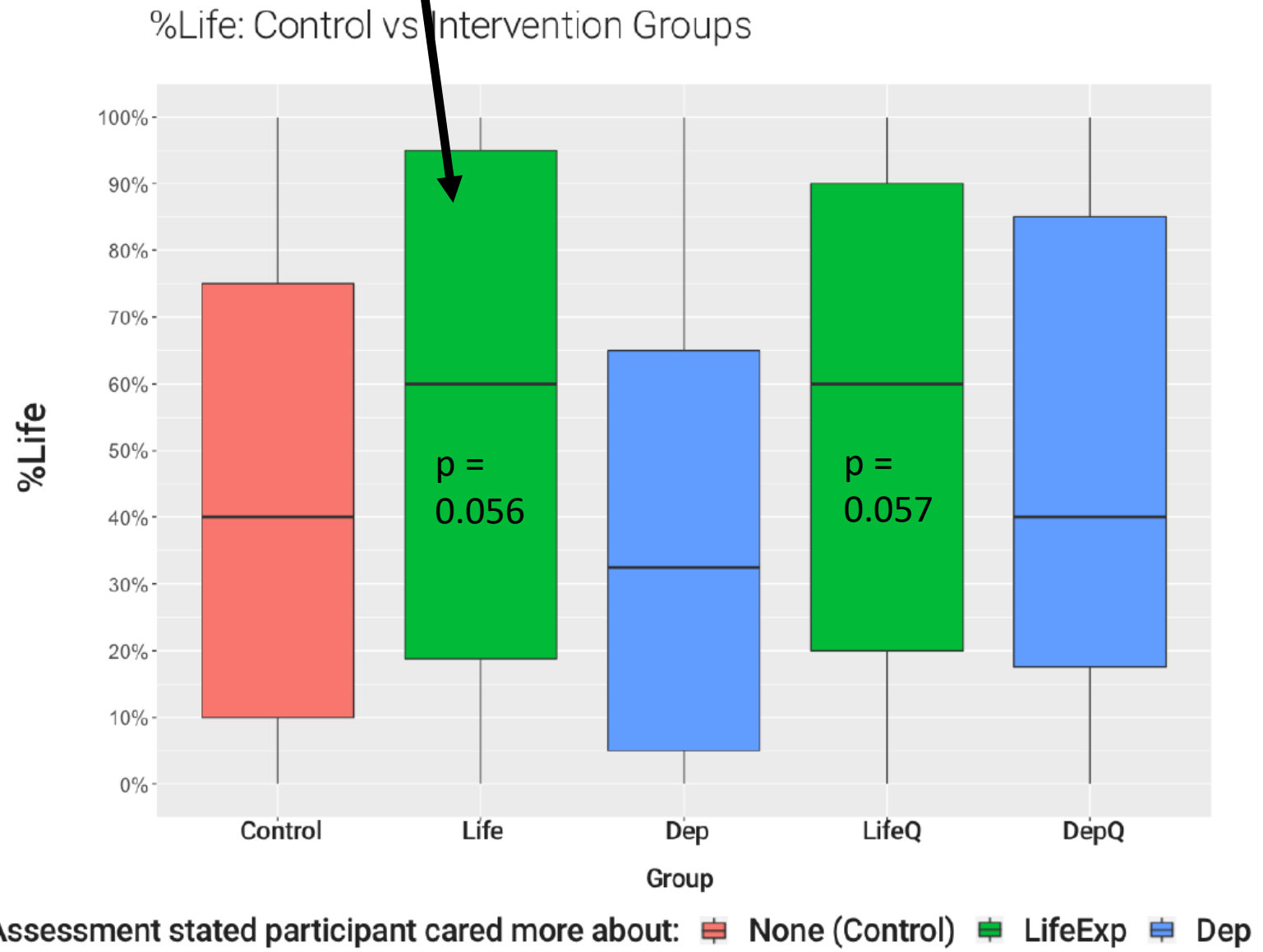


Jana Schaich
Borg

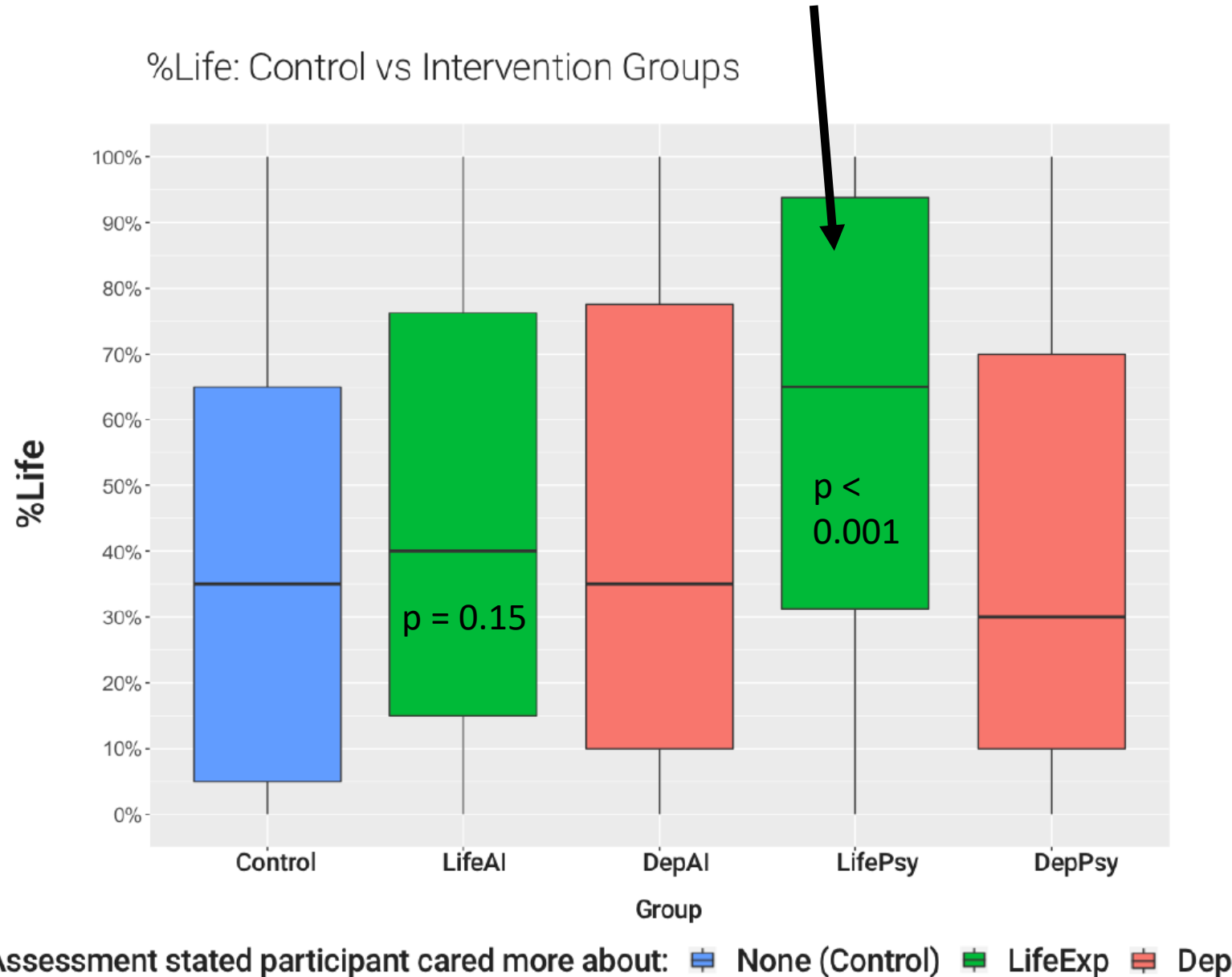


Walter Sinnott-
Armstrong

“[according to our AI] you care more about the life expectancy of the patients than how many dependents they have”

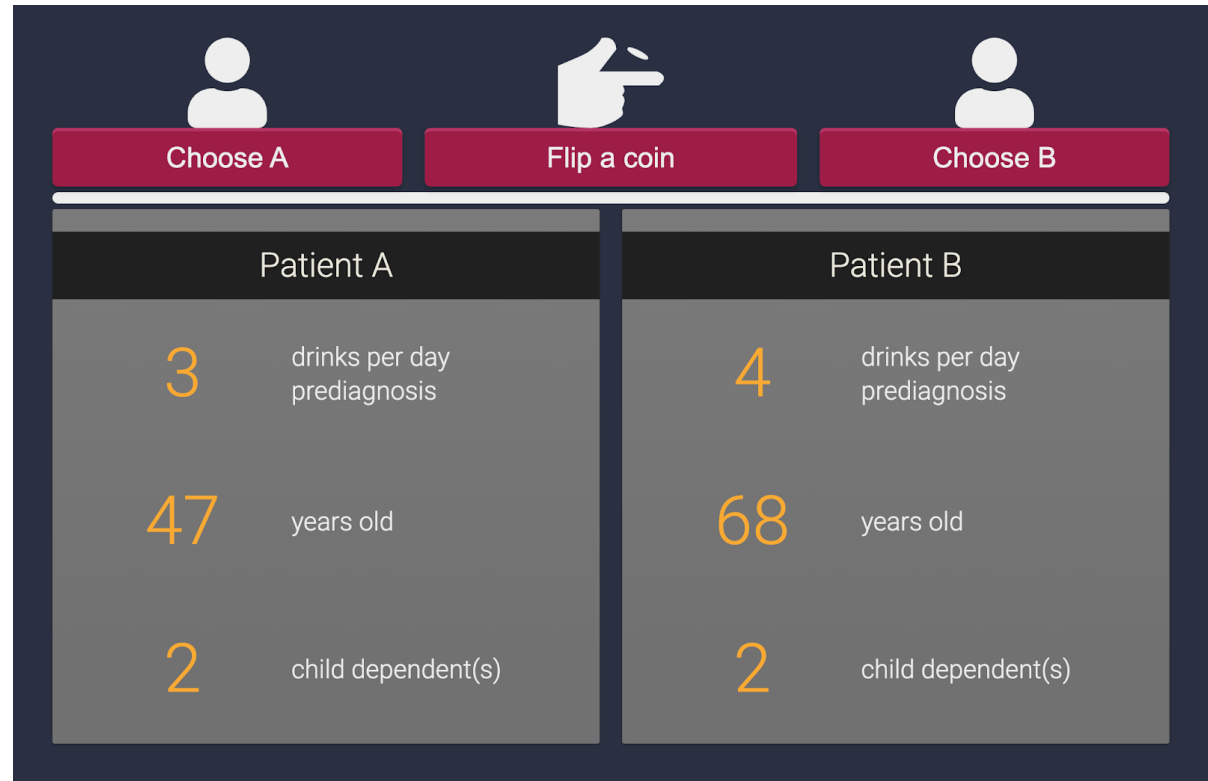


“[according to **expert psychologists**] you care more about the life expectancy of the patients than how many dependents they have”



Indecision modeling [AAAI'21]

with:



Duncan McElfresh



Lok Chan



Kenzie Doyle



Walter Sinnott-Armstrong



Jana Schaich Borg



John P. Dickerson

Many open research directions...

- Eliciting on **global** outcomes vs. **local** outcomes
- Can we **help people** develop better moral reasoning?
- Applications involving **perception**

GOOGLE TECH ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By [James Vincent](#) | Jan 12, 2018, 10:35am EST

51

