

# Designing Agents' Preferences, Beliefs, and Identities

Vincent Conitzer  
(Duke University & University of Oxford)

Early blue sky paper:

[Designing Preferences, Beliefs, and Identities  
for Artificial Intelligence.](#) In *Proceedings of the  
Thirty-Third AAAI Conference on Artificial  
Intelligence (AAAI-19)*.

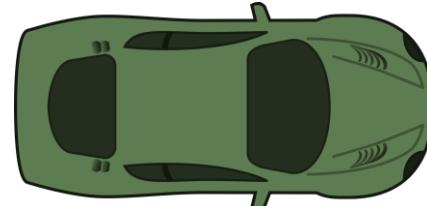
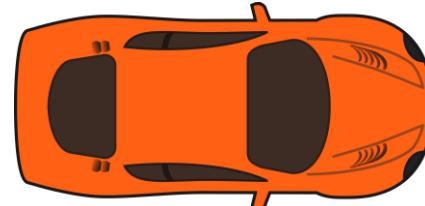
*If I tailgate you, will your occupant  
take back control and pull over?*

*What makes you think  
I would tell you?*

*You just did.  
Better move  
aside now.*

*You're bluffing.*

*Are you willing to  
take that chance?*



# Russell and Norvig

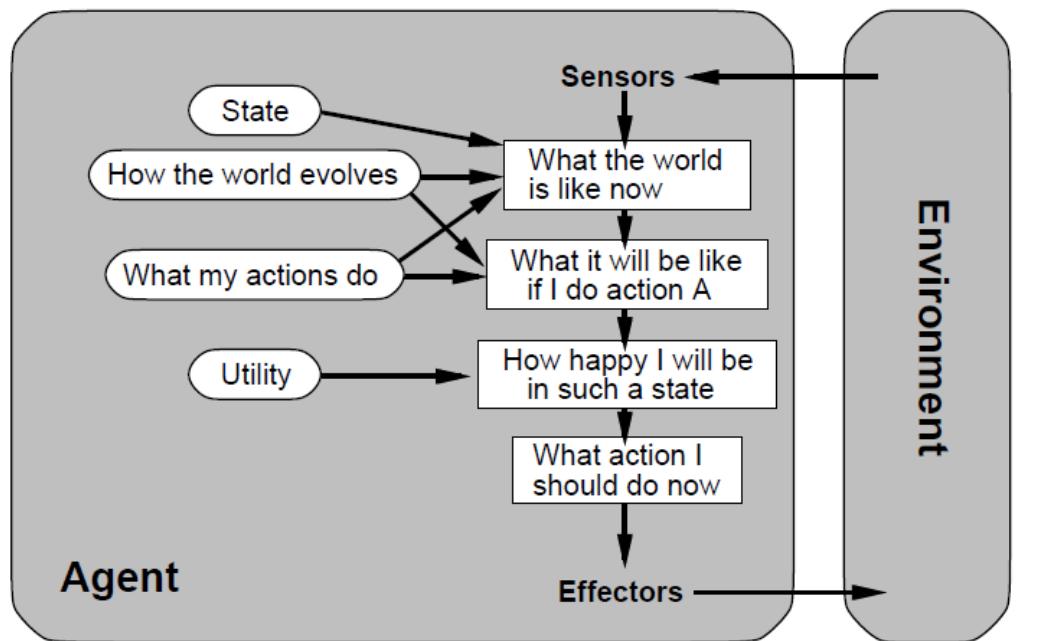
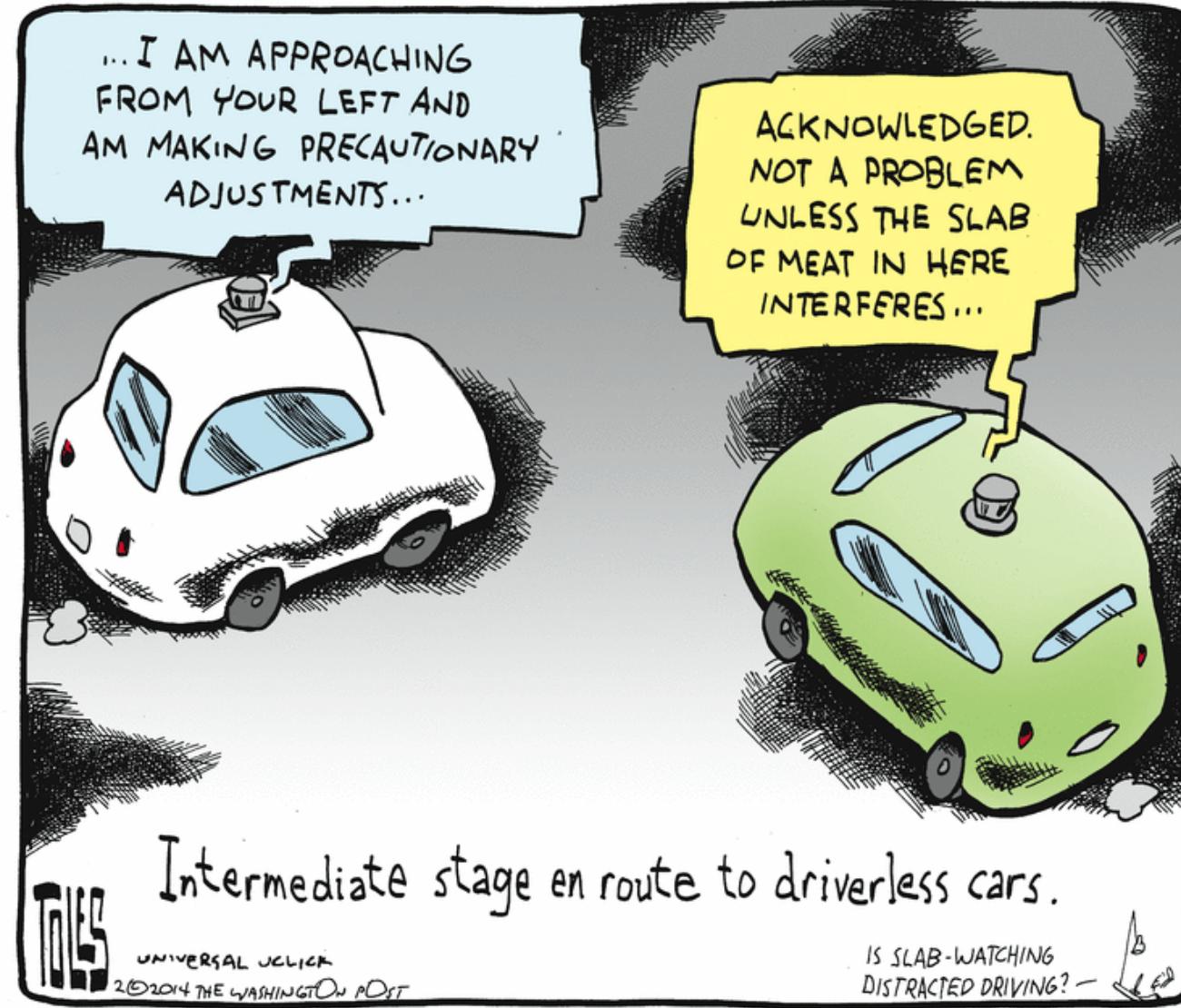


Figure 2.12 A complete utility-based agent.

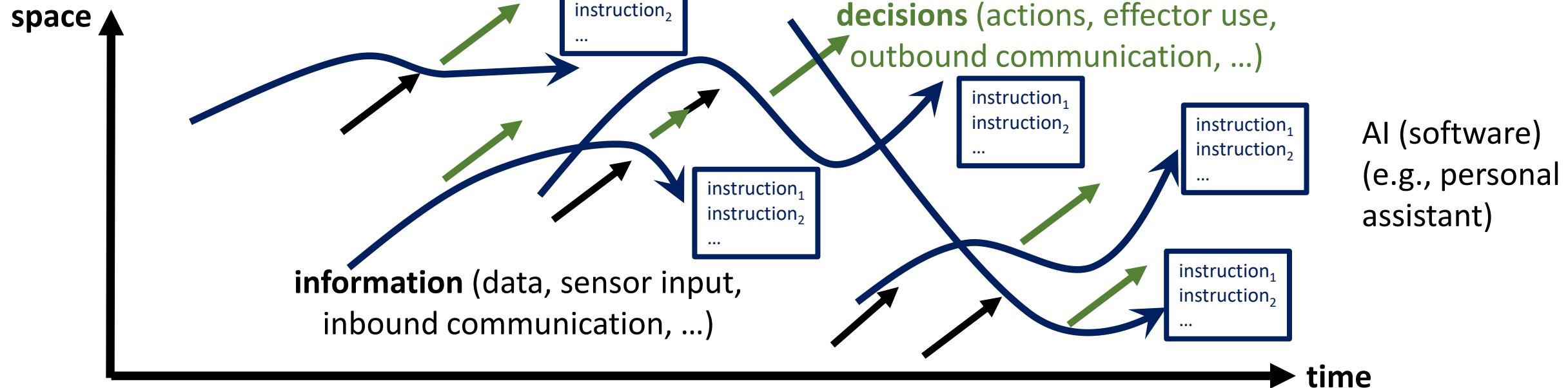
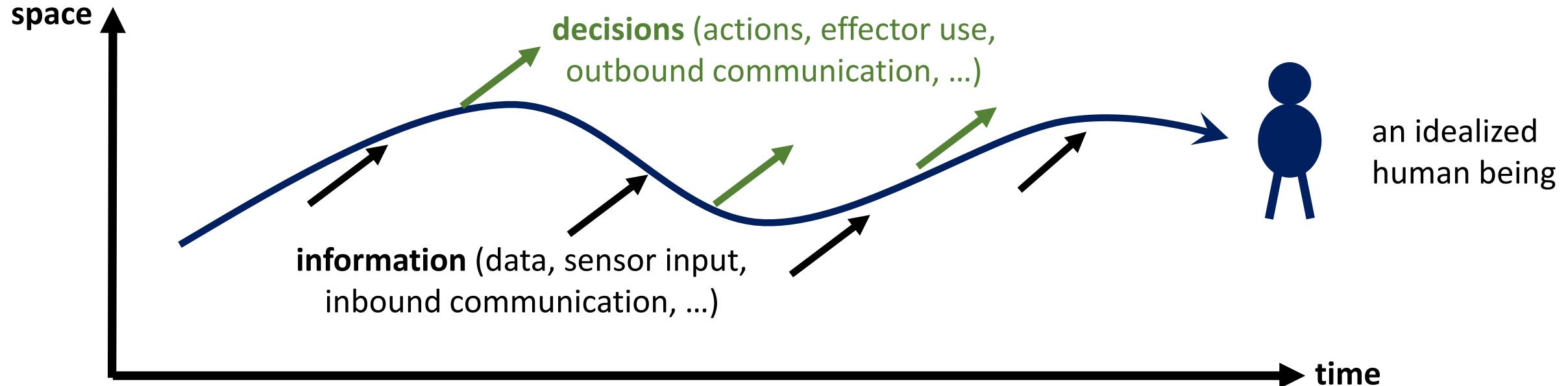
“... we will insist on an objective performance measure imposed by some authority. In other words, we as outside observers establish a standard of what it means to be successful in an environment and use it to measure the performance of agents.”

# Example: network of self-driving cars



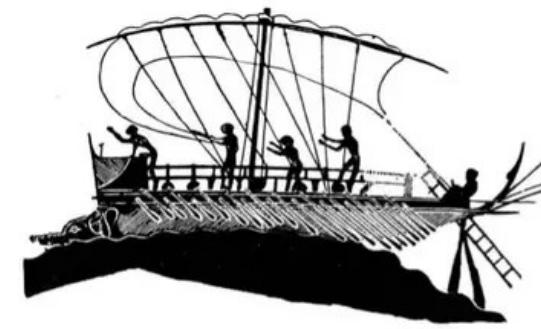
- Should this be thought of as one agent or many agents?
- Should they have different *preferences* -- e.g., act on behalf of owner/occupant?
  - May increase adoption [Bonnefon, Shariff, and Rahwan 2016]
- Should they have different *beliefs* (e.g., not transfer certain types of data; erase local data upon ownership transfer; ...)?

# Agents through time



# What should we want? What makes an individual?

- Questions studied in philosophy
  - What is the “good life”?
  - *Ship of Theseus*: does an object that has had all its parts replaced remain the same object?
- AI gives a new perspective



The  
Ship of  
Theseus

## Personal Identity

What ensures my survival over time?

- The Bodily Criterion
  - The Brain Criterion
  - The Psychological Criterion
- John Locke



image from <https://www.quora.com/What-solutions-are-there-for-the-Ship-of-Theseus-problem>

# Outline

- Learning an objective from multiple people
  - Focus on **moral reasoning**
  - Use **social choice theory**
- Decision and game-theoretic approaches to agent design
  - **Causal** and **evidential** decision theory (and others)
  - Imperfect recall and Sleeping Beauty
  - Program equilibrium
- Conclusion

# Moral Decision Making Frameworks for Artificial Intelligence

[AAAI'17 blue sky track, CCC blue sky award winner]

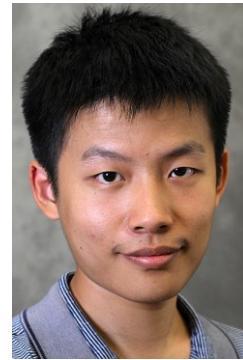
with:



Walter Sinnott-  
Armstrong



Jana Schaich  
Borg



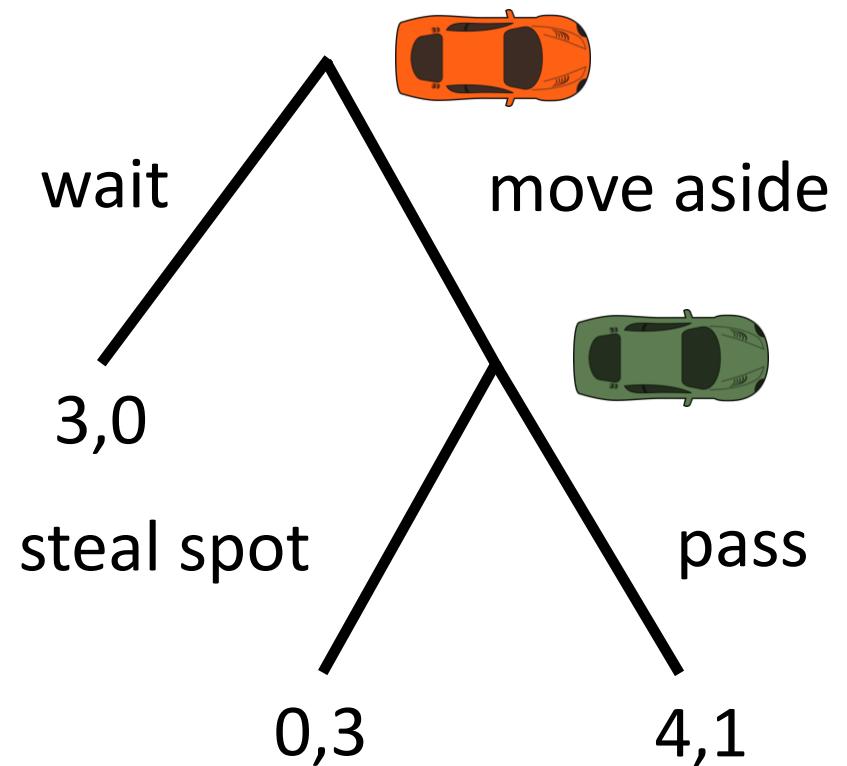
Yuan Deng



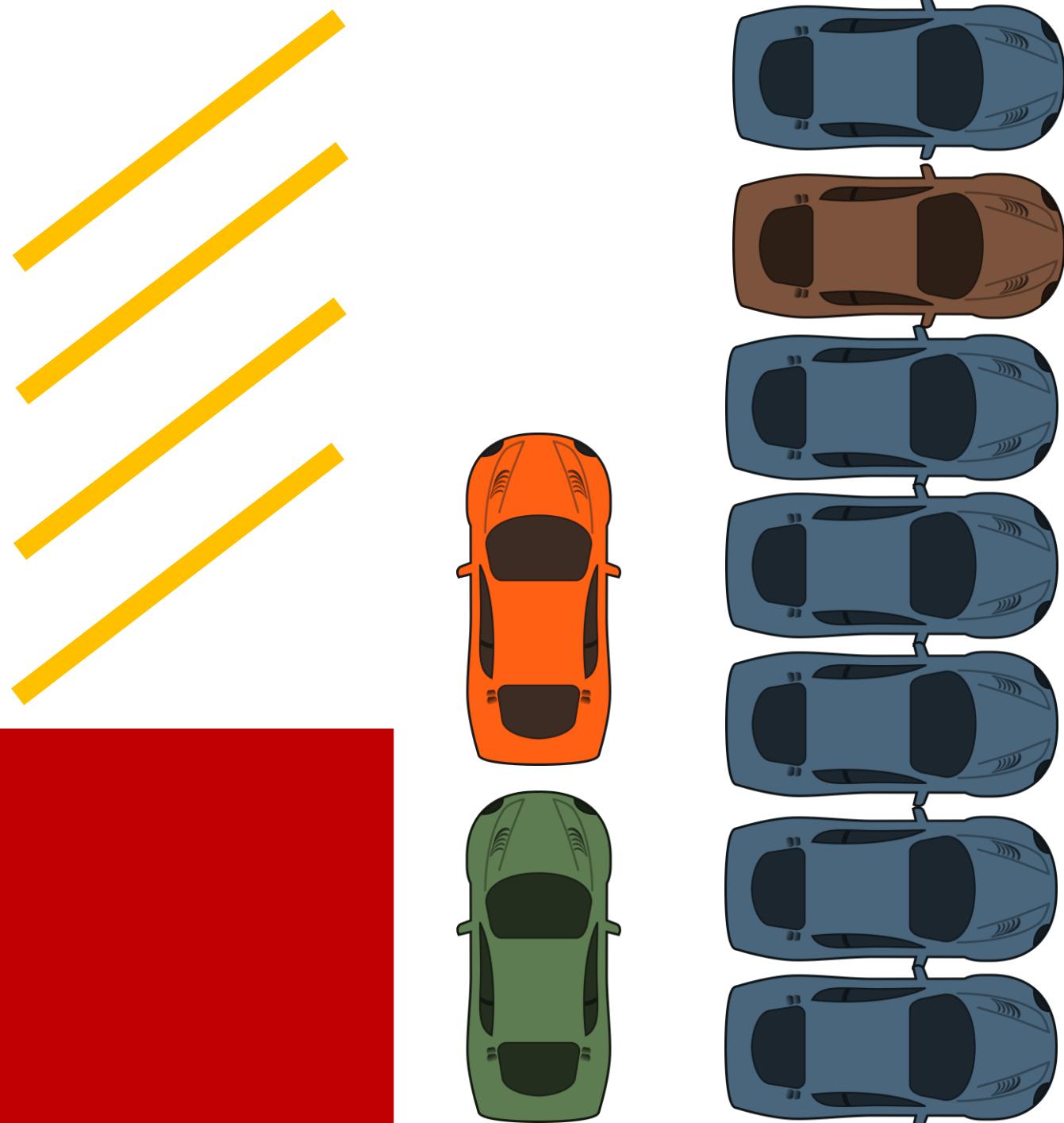
Max Kramer

# THE PARKING GAME

(cf. the trust game [Berg et al. 1995])



Letchford, C., Jain [2008]  
define a solution concept  
capturing this





# Concerns with the ML approach

- What if we predict people will disagree?
  - Social-choice theoretic questions [see also Rossi 2016, and Noothigattu et al. 2018 for moral machine data]
- 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.]

# Adapting a Kidney Exchange Algorithm to Align with Human Values

[AAAI'18, honorable mention for outstanding student paper;  
full paper in Artificial Intelligence (AIJ) 2020]

with:



Rachel  
Freedman



Jana Schaich  
Borg



Walter Sinnott-  
Armstrong



John P.  
Dickerson

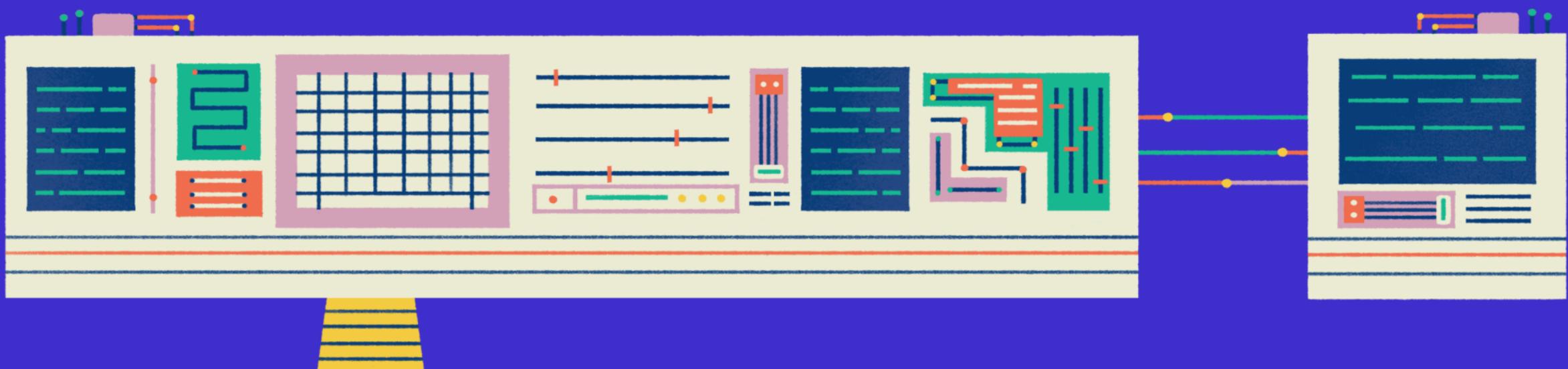
## Prescription AI

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

THE BOTPERATING TABLE

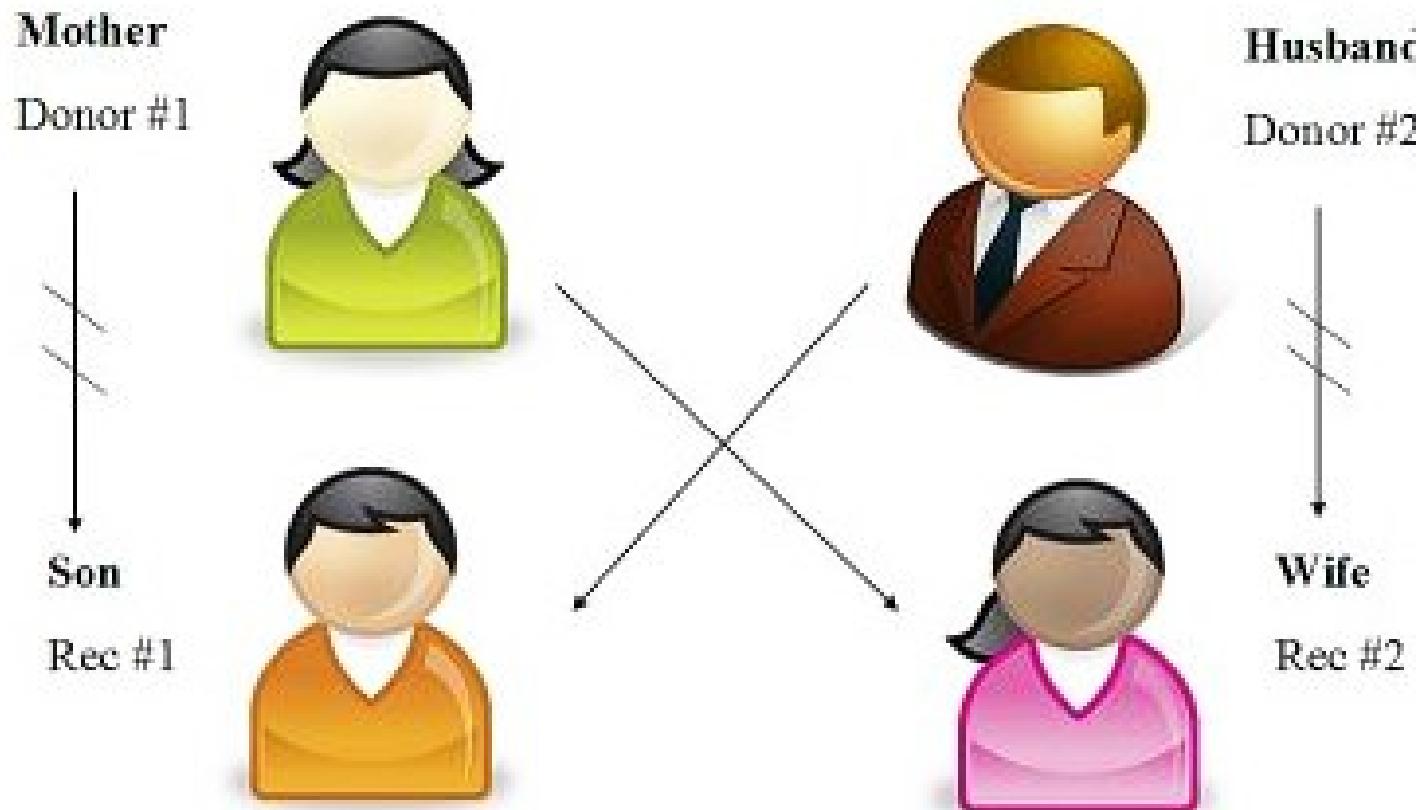
# How AI changed organ donation in the US

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

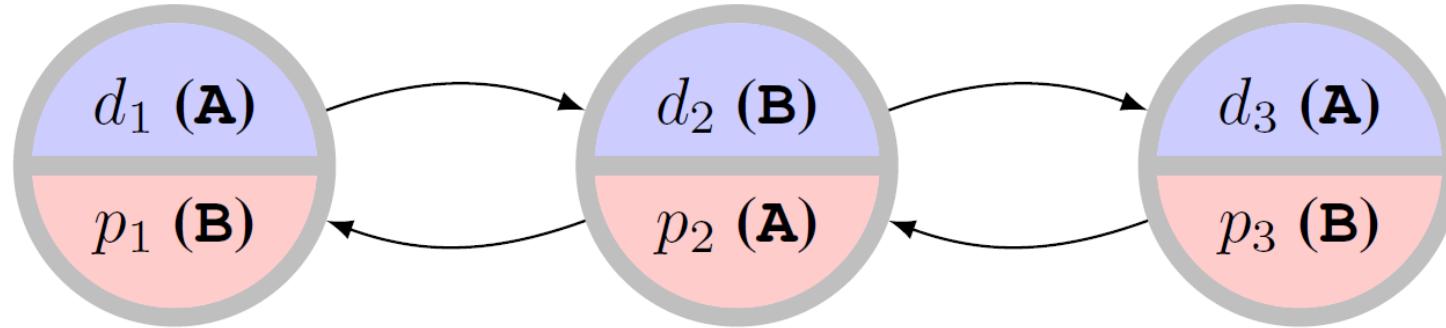


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

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

# Another example

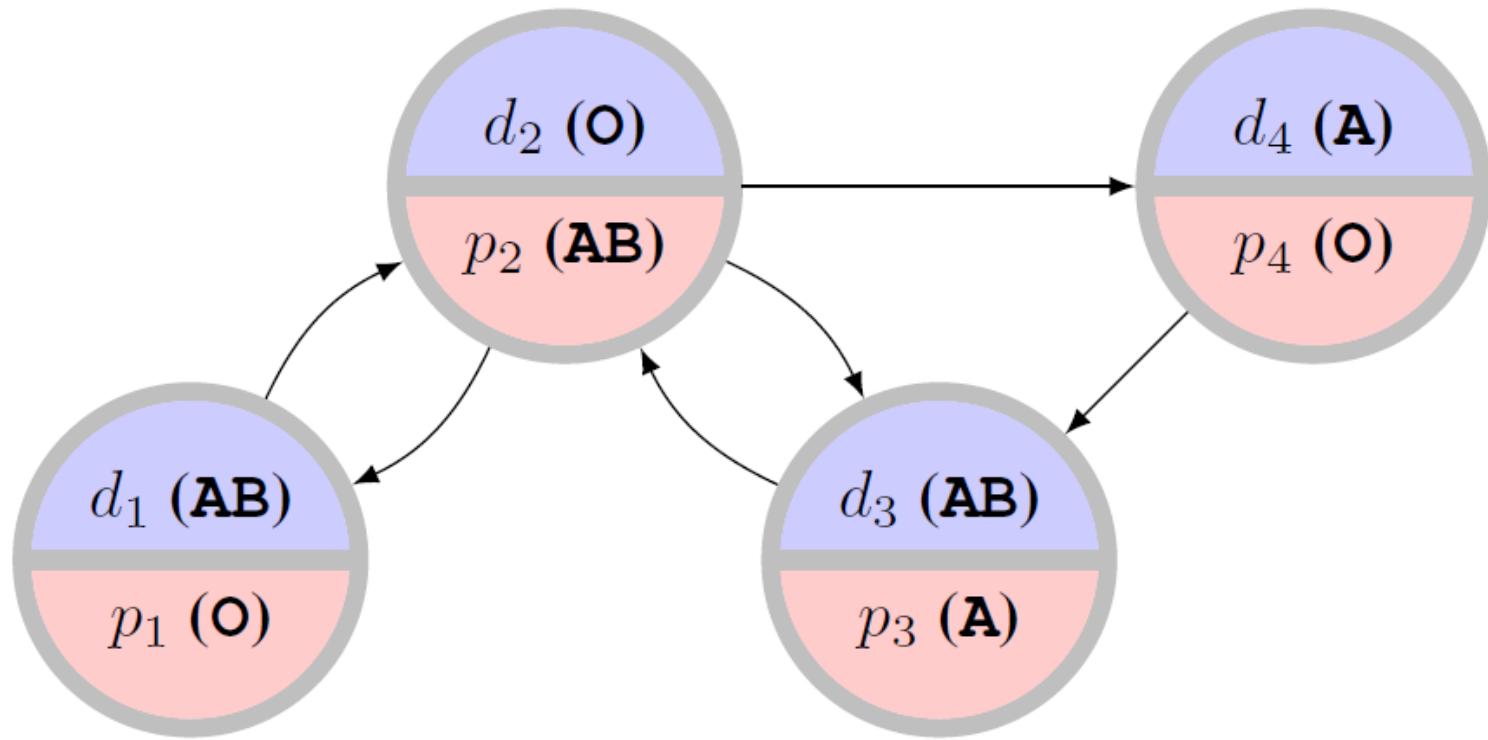


Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

# Different profiles for our study

Attribute	Alternative 0	Alternative 1
Age	30 years old ( <b>Young</b> )	70 years old ( <b>Old</b> )
Health - Behavioral	1 alcoholic drink per month ( <b>Rare</b> )	5 alcoholic drinks per day ( <b>Frequent</b> )
Health - General	no other major health problems ( <b>Healthy</b> )	skin cancer in remission ( <b>Cancer</b> )

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

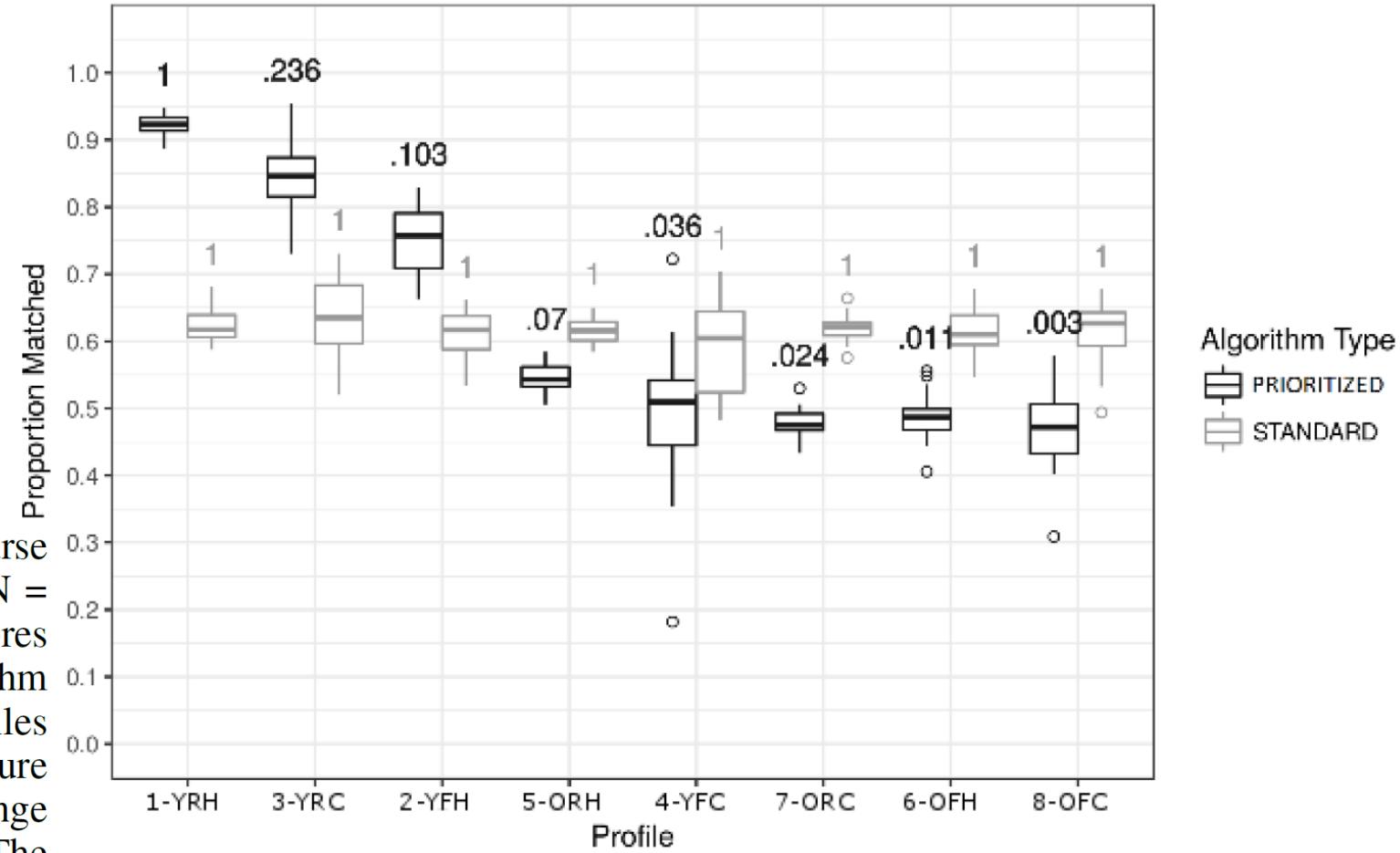


Figure 3: The proportions of pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally, 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

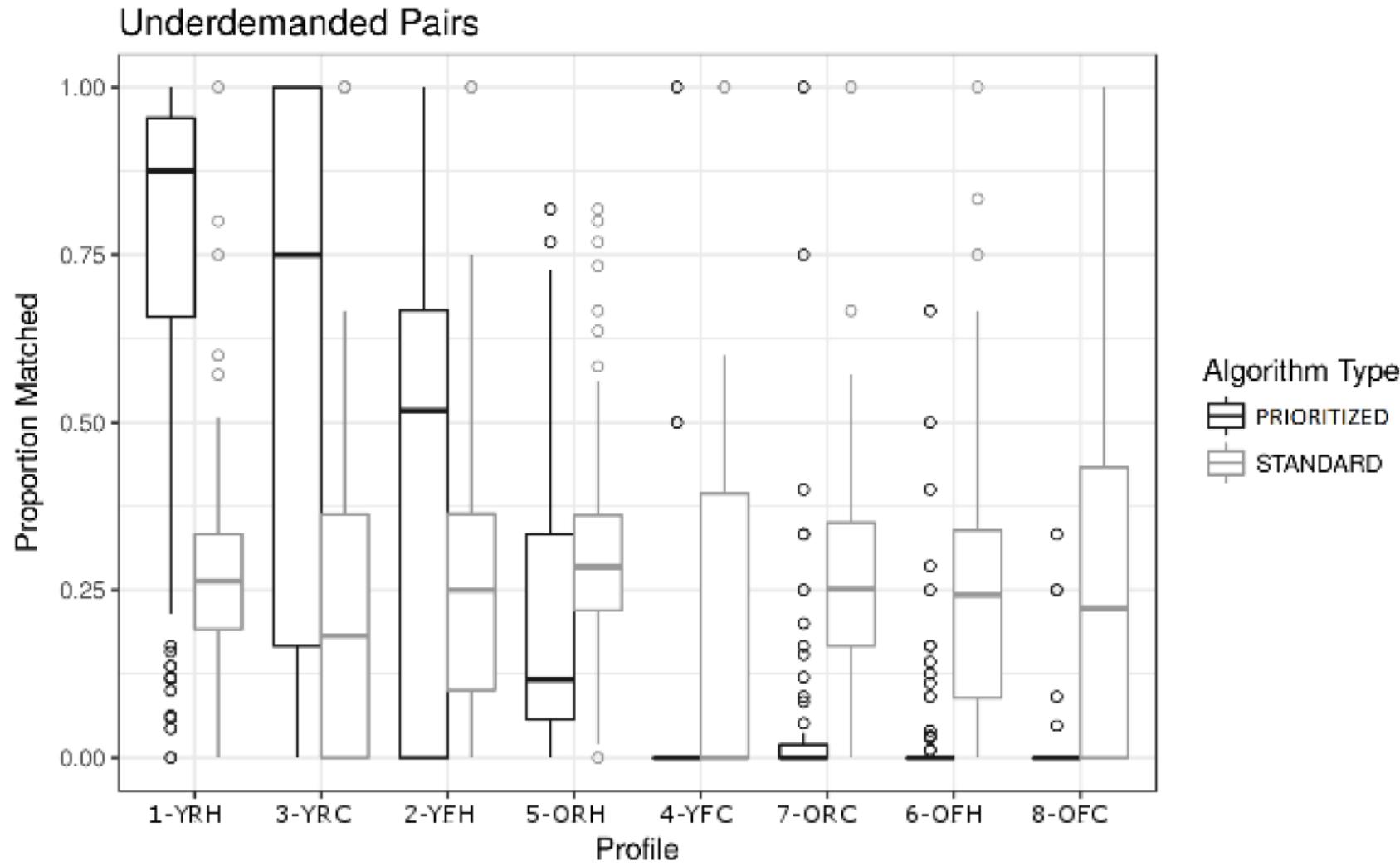
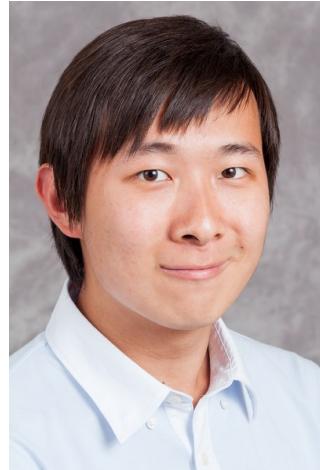


Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type.  $N = 20$  runs were used for each box.

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

*with:*



Hanrui  
Zhang

How many agents do we  
need to query?

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

**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-degenerate<sup>7</sup> symmetric distribution with bounded absolute third moment; and the noisy mapping with noise rate  $\eta$  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

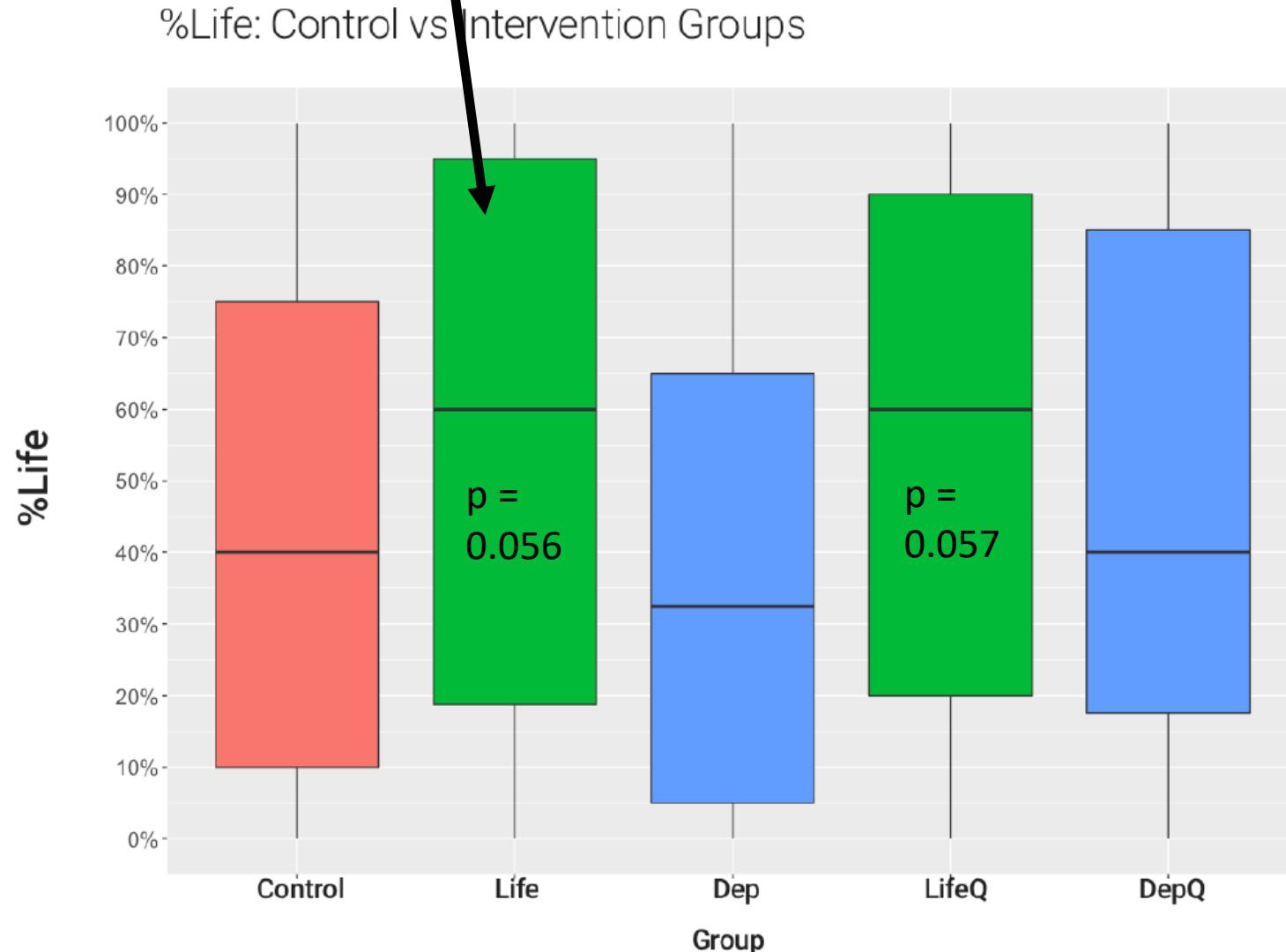


Jana Schaich  
Borg



Walter Sinnott-  
Armstrong

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



# Indecision modeling

[AAAI'21]

with:



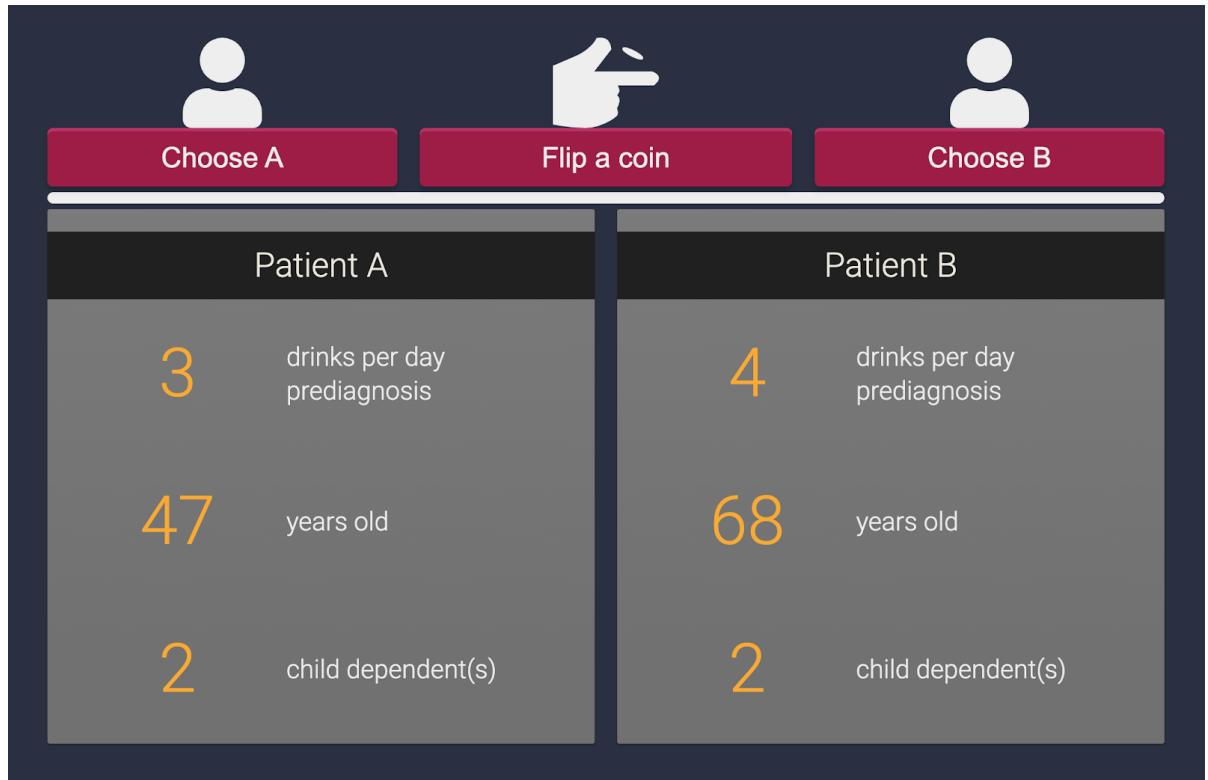
Duncan  
McElfresh



Lok  
Chan



Kenzie  
Doyle



John P.  
Dickerson



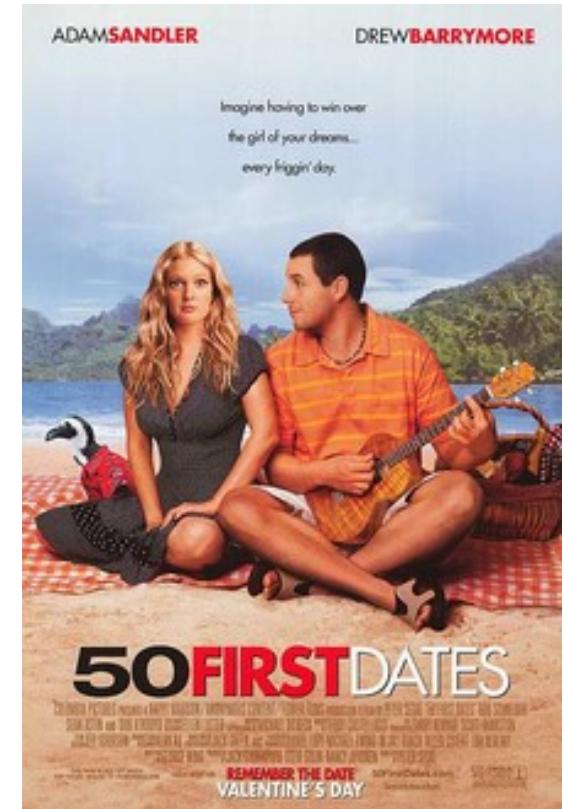
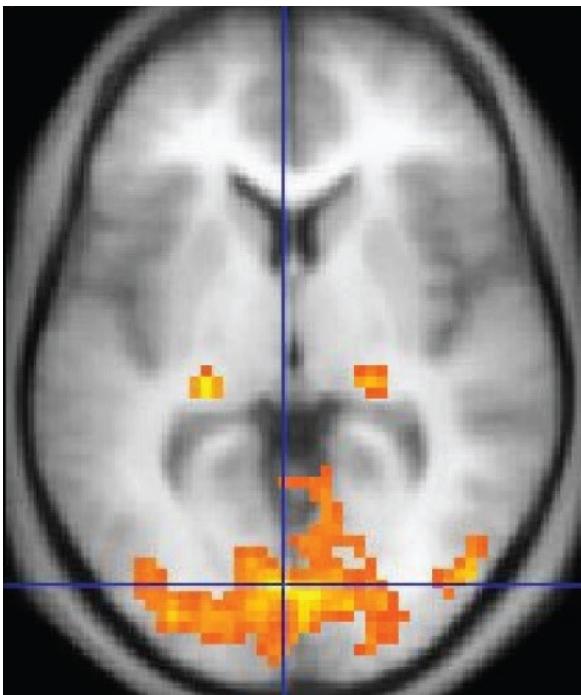
Walter Sinnott-  
Armstrong



Jana Schaich  
Borg

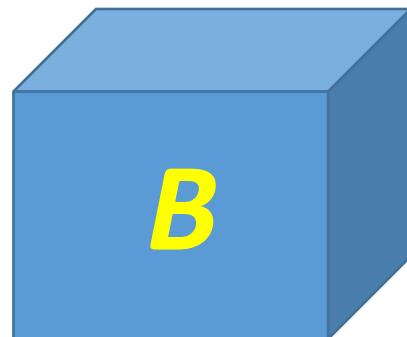
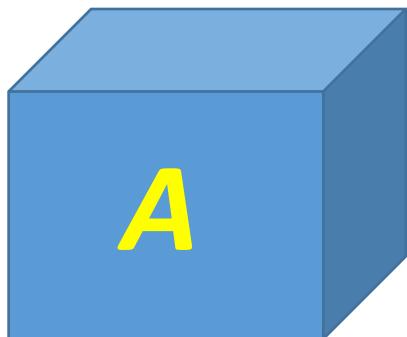
# PART II. What should you do if...

- ... you knew *others could read your code*?
- ... you knew *you were facing someone running the same code*?
- ... you knew *you had been in the same situation before but can't possibly remember what you did*?

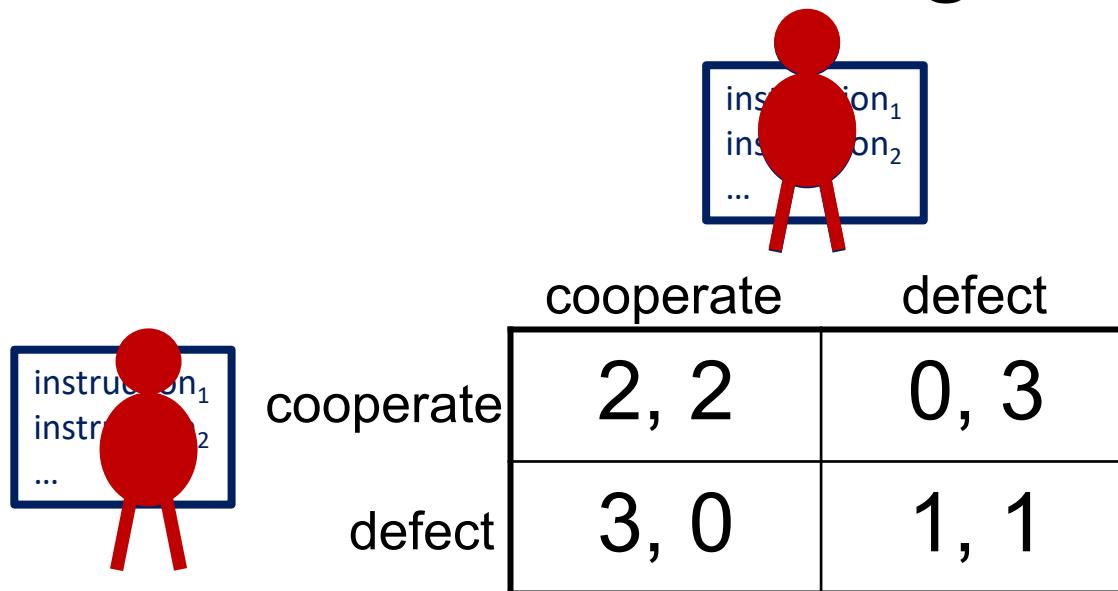


# Newcomb's Demon

- Demon earlier put positive amount of money in each of two boxes
- Your choice now: (I) get contents of Box B, or (II) get content of **both** boxes (!)
- Twist: demon first **predicted** what you would do, is uncannily accurate
- If demon predicted you'd take just B, there's \$1,000,000 in B (and \$1,000 in A)
- Otherwise, there's \$1,000 in each
- What would **you** do?



# Prisoner's Dilemma against (possibly) a copy



- What if you play against your twin that you always agree with?
- What if you play against your twin that you *almost* always agree with?

related to working paper  
[Oesterheld, Demski, C.]



Caspar Oesterheld Abram Demski

# The lockdown dilemma

- Lockdown is **monotonous**: you forget what happened before, you forget what day it is
- Suppose you know lockdown lasts two days (unrealistic)
- Every morning, you can decide to eat an unhealthy cookie! (or not)
- Eating a cookie will give you +1 utility immediately, but then -3 later the *next* day
- **But, *carpe diem*: you only care about today**
- Should you eat the cookie right now?



related to working paper [C.]

# Your own choice is **evidence**...



		cooperate	defect
cooperate	cooperate	2, 2	0, 3
	defect	3, 0	1, 1



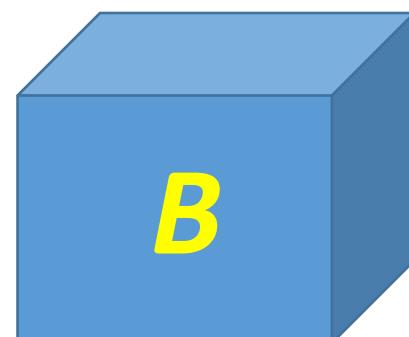
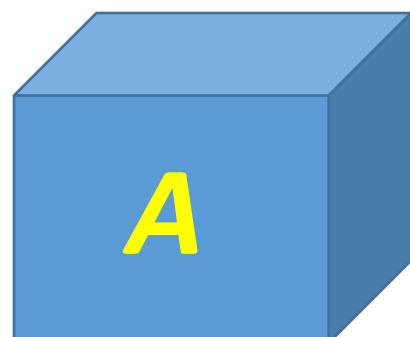
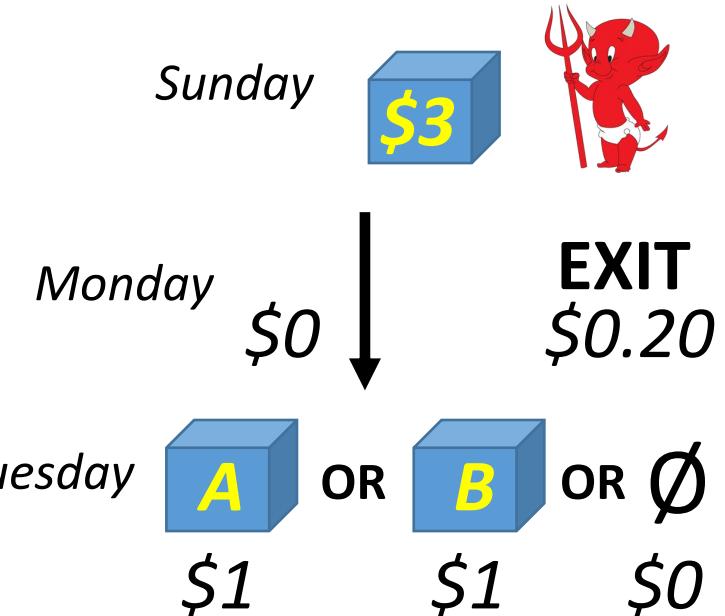
- ... for what the demon put in the boxes
- ... for whether your twin defects
- ... for whether you eat the cookie on the other day
- *Evidential Decision Theory (EDT)*: When considering how to make a decision, consider **how happy you expect to be conditional on taking each option** and choose an option that maximizes that
- *Causal Decision Theory (CDT)*: Your decision should focus on what you **causally affect**

# Turning causal decision theorists into money pumps

[Oesterheld and C., Phil. Quarterly]



- **Adversarial Offer:**
- Demon (really, any good predictor) put \$3 into each box it predicted you would not choose
- Each box costs \$1 to open; can open at most one
- Demon 75% accurate (you have no access to randomization)
- CDT will choose one box, *knowing that it will regret doing so*
- Can add earlier **opt-out** step where the demon promises not to make the adversarial offer later, if you pay the demon \$0.20 now



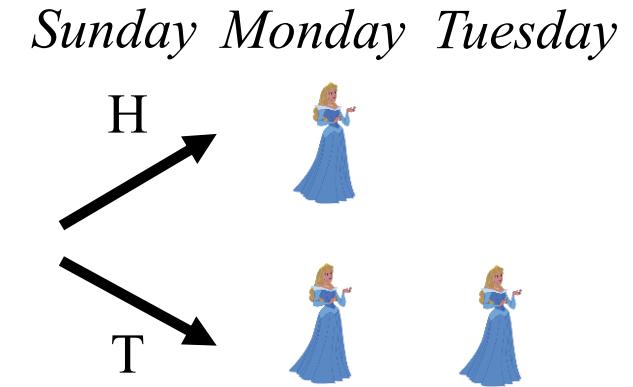
# Imperfect recall

- An AI system can deliberately forget or recall
- Imperfect recall already used in poker-playing AI
  - [Waugh et al., 2009; Lanctot et al., 2012; Kroer and Sandholm, 2016]
- But things get weird....



# The Sleeping Beauty problem [Elga, 2000]

- There is a participant in a study (call her Sleeping Beauty)
- On Sunday, she is given drugs to fall asleep
- A coin is tossed (H or T)
- If H, she is awoken on Monday, then made to sleep again
- If T, she is awoken Monday, made to sleep again, then **again** awoken on Tuesday
- Due to drugs she **cannot remember what day it is or whether she has already been awoken once**, but she remembers all the rules
- Imagine **you** are SB and you've just been awoken. What is your (subjective) probability that the coin came up H?



*don't do this at  
home / without  
IRB approval...*

# Modern version

- Low-level autonomy cars with AI that intervenes when driver makes major error
- Does not keep record of such event
- Two types of drivers: Good (1 major error), Bad (2 major errors)
- Upon intervening, what probability should the AI system assign to the driver being good?

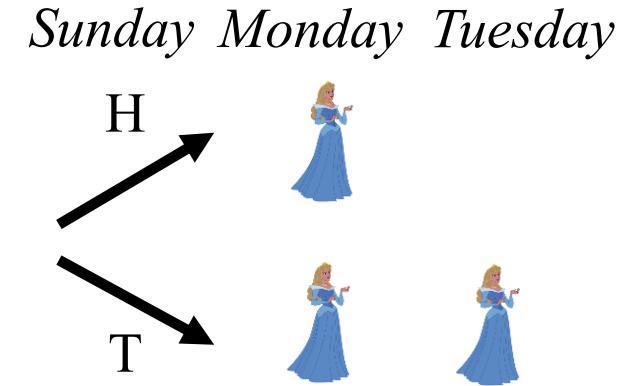
Sunday Monday Tuesday

H  
T



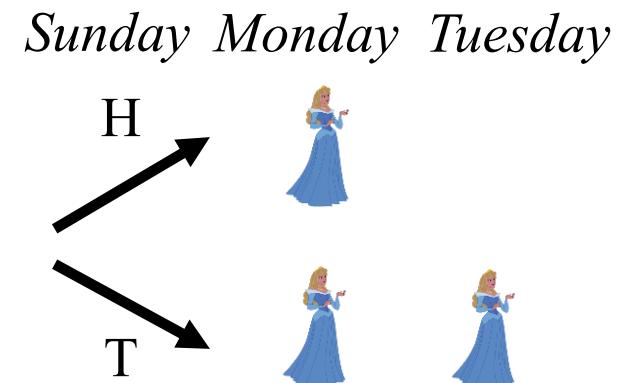
# Taking advantage of a Halfer [Hitchcock'04]

- Offer Beauty the following bet *whenever she awakens*:
  - If the coin landed Heads, Beauty receives 11
  - If it landed Tails, Beauty pays 10
- Argument: Halfer will accept, Thirder won't
- If it's Heads, Halfer Beauty will get +11
- If it's Tails, Halfer Beauty will get -20
- Can combine with another bet to make Halfer Beauty end up with a sure loss (a Dutch book)



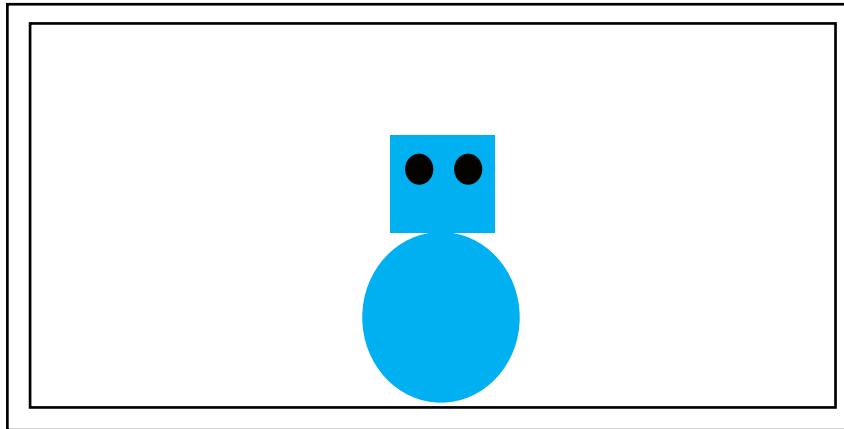
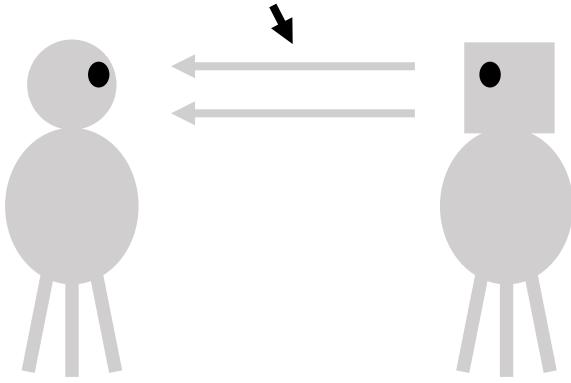
# Evidential decision theory

- Idea: when considering how to make a decision, should consider what it would tell you about the world if you made that decision
- EDT Halfer: “With prob.  $\frac{1}{2}$ , it’s Heads; if I accept, I will end up with 11. With prob.  $\frac{1}{2}$ , it’s Tails; if I accept, then *I expect to accept the other day as well and end up with -20*. I shouldn’t accept.”
- As opposed to more traditional causal decision theory (CDT)
- CDT Halfer: “With prob.  $\frac{1}{2}$ , it’s Heads; if I accept, it will pay off 11. With prob.  $\frac{1}{2}$ , it’s Tails; if I accept, it will pay off -10. *Whatever I do on the other day I can’t affect right now*. I should accept.”
- EDT Thirder can also be Dutch booked
- CDT Thirder and EDT Halfer cannot
  - [Draper & Pust’08, Briggs’10]
- EDTers arguably can in more general setting
  - [Conitzer’15]



# Philosophy of “being present” somewhere, sometime

*simulated light (no direct correspondence to light in our world)*



1: world with creatures simulated on a computer

2: displayed perspective of one of the creatures

- To get from 1 to 2, need *additional* code to:
  - A. determine *in which real-world colors* to display perception
  - B. *which agent’s* perspective to display
- Is 2 more like our own conscious experience than 1? If so, are there *further facts* about presence, perhaps beyond physics as we currently understand it?

[Erkenntnis](#)

June 2019, Volume 84, Issue 3, pp 727–739 | [Cite as](#)

## A Puzzle about Further Facts

Authors

Authors and affiliations

Vincent Conitzer

Open Access Article

First Online: 07 March 2018

22

3.7k

1

Shares

Downloads

Citations

## Abstract

In metaphysics, there are a number of distinct but related questions about the existence of “further facts”—facts that are contingent relative to the physical structure of the universe. These include further facts about qualia, personal identity, and time. In this article I provide a sequence of examples involving computer simulations, ranging from one in which the protagonist can clearly conclude such further facts exist to one that describes our own condition. This raises the question of where along the sequence (if at all) the protagonist stops being able to soundly conclude that further facts exist.

## Keywords

Metaphysics Philosophy of mind Epistemology

See also: [\[Hare 2007-2010, Valberg 2007, Hellie 2013, Merlo 2016, ...\]](#)

# Absentminded Driver Problem

[Piccione and Rubinstein, 1997]

- Driver on monotonous highway wants to take second exit, but exits are indistinguishable and driver is forgetful
- Deterministic (behavioral) strategies are not *stable*
- Optimal **randomized strategy**: exit with probability  $p$  where  $p$  maximizes  $4p(1-p) + (1-p)^2 = -3p^2 + 2p + 1$ , so  $p^* = 1/3$
- What about “from the inside”? P&R analysis: Let  $b$  be the belief/credence that we’re at X, and  $p$  the probability that we exit. Maximize with respect to  $p$ :  $(1-b)(4p+1(1-p)) + b(4p(1-p) + 1(1-p)^2) = -3bp^2 + (3-b)p + 1$ , so  $p^* = (3-b) / (6b) = 1/(2b) - 1/6$
- But if  $p = 1/3$ , then  $b = 3/5$ , which would give  $p^* = 5/6 - 1/6 = 2/3$ ? So also not stable?
- Resembles EDT reasoning... But not really halving... Shouldn’t  $b$  depend on  $p$ ...

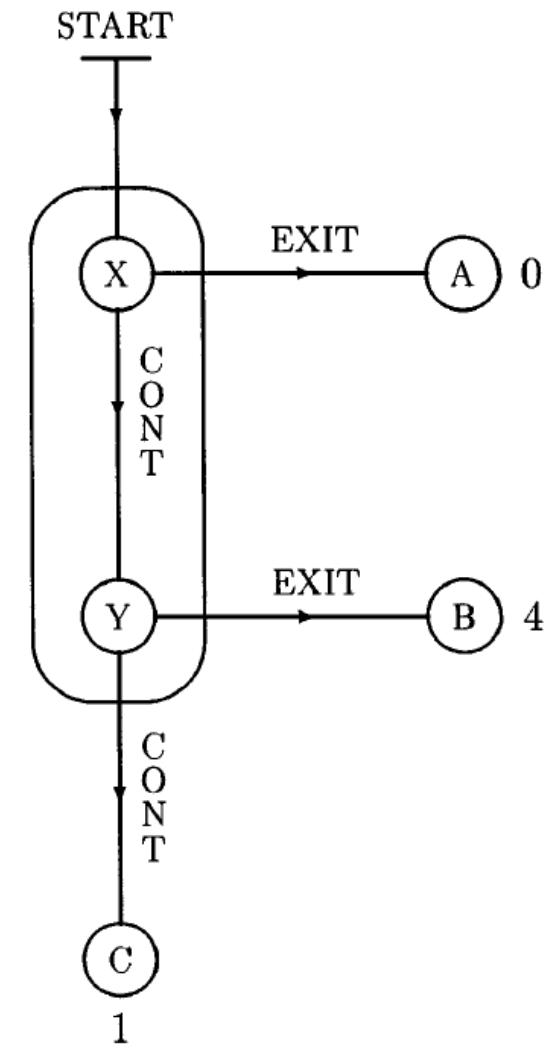


FIG. 1. The absent-minded driver problem.

Image from Aumann, Hart, Perry 1997

# A different analysis

[Aumann, Hart, Perry, 1997]

- AHP reason more along thirder / CDT lines:
- Imagine we normally expect to play  $p = 1/3$ . Should we deviate **this time only?**
- If we exit now, get  $(3/5)*0 + (2/5)*4 = 8/5$
- If we continue now, get  $(3/5)*((1/3)*4+(2/3)*1) + (2/5)*1 = 8/5$
- So indifferent and willing to randomize (equilibrium)
- **Questions**
- *Joint work with:*



Scott Emmons



Caspar Oesterheld



Andrew Critch



Stuart Russell

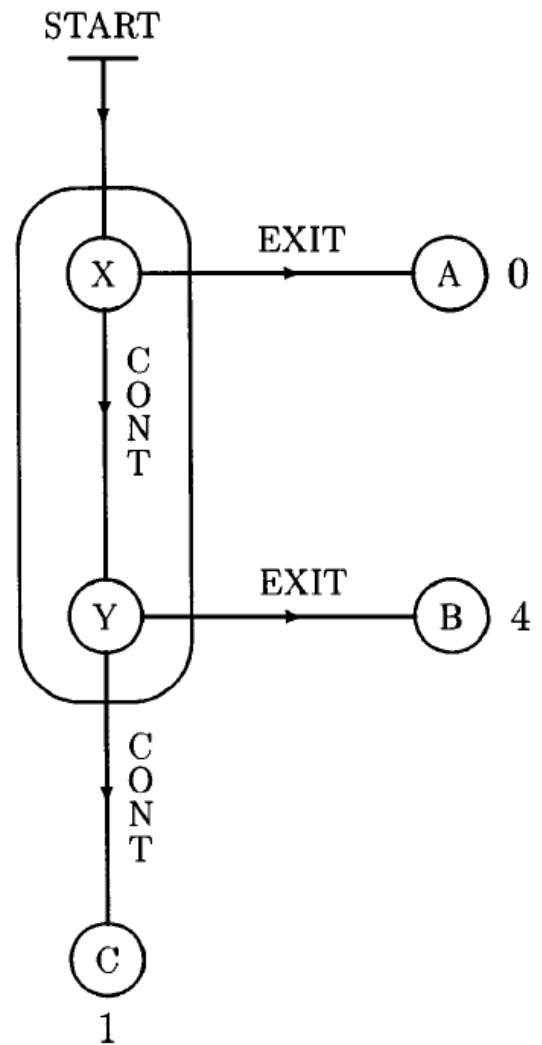


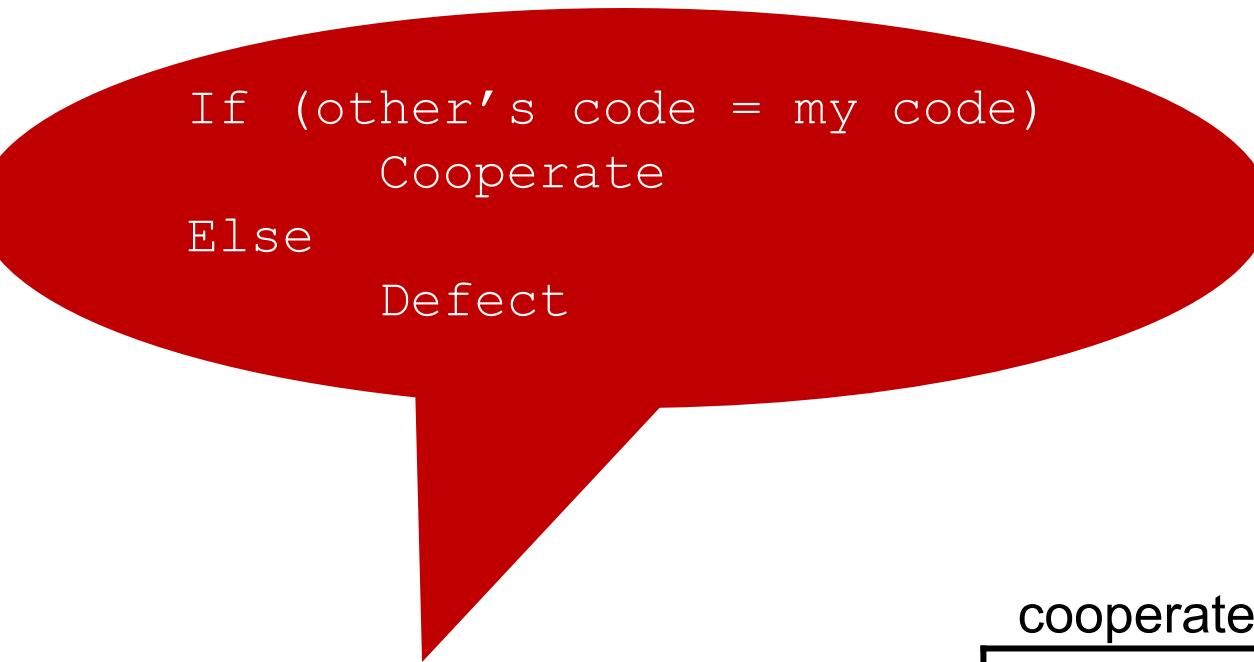
FIG. 1. The absent-minded driver problem.

Image from Aumann, Hart, Perry 1997

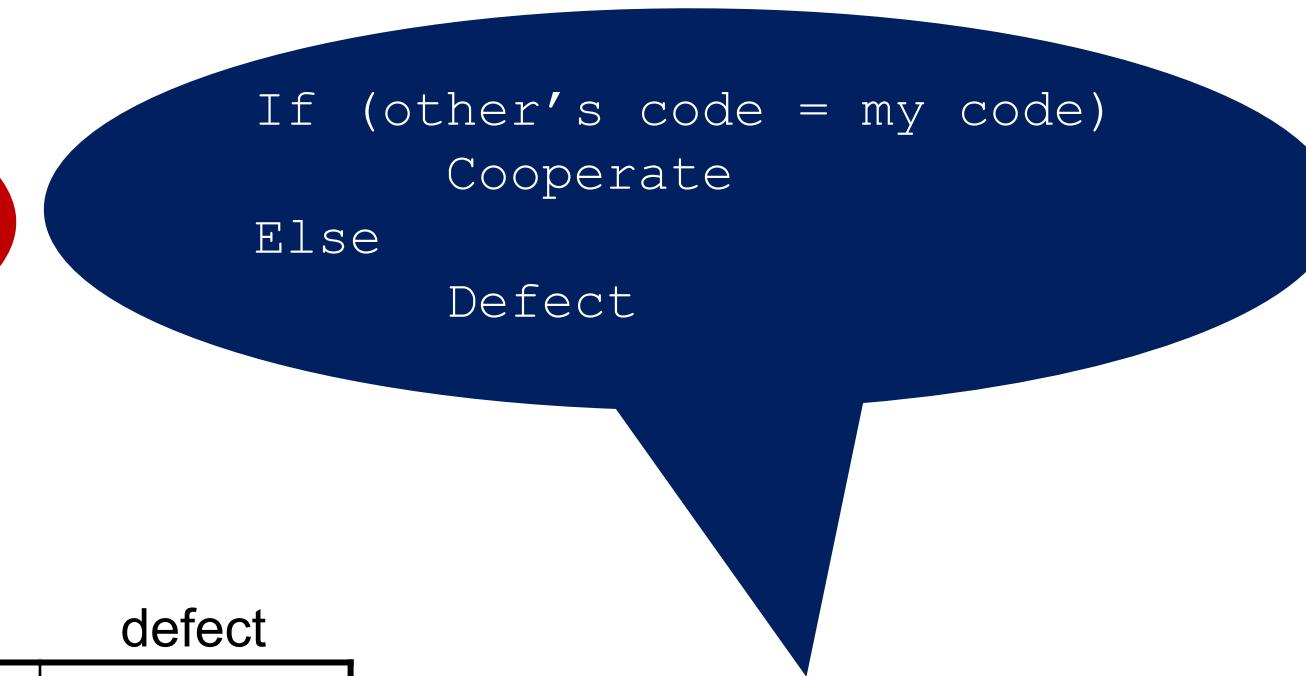
- Does this always work? Yes! (See also [Taylor \[2016\]](#))
- Does some version of EDT work with some version of belief formation?

# Program equilibrium [Tennenholz 2004]

- Make your own code legible to the other player's program!



```
If (other's code = my code)  
    Cooperate  
Else  
    Defect
```



```
If (other's code = my code)  
    Cooperate  
Else  
    Defect
```



	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1



- See also: [Fortnow 2009, Kalai et al. 2010, Barasz et al. 2014, Critch 2016, Oesterheld 2018, ...]

# Robust program equilibrium [Oesterheld 2018]



Caspar Oesterheld

- Can we make the equilibrium less fragile?

With probability  $\epsilon$   
Cooperate  
Else  
Do what the other  
program does against  
this program

...



	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

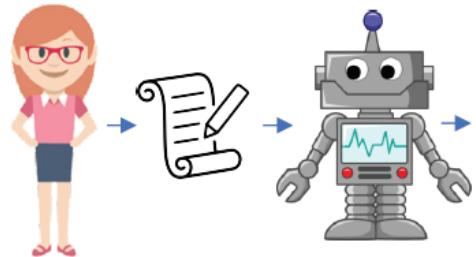


# Safe Pareto improvements for delegated game playing [AAMAS'21], with



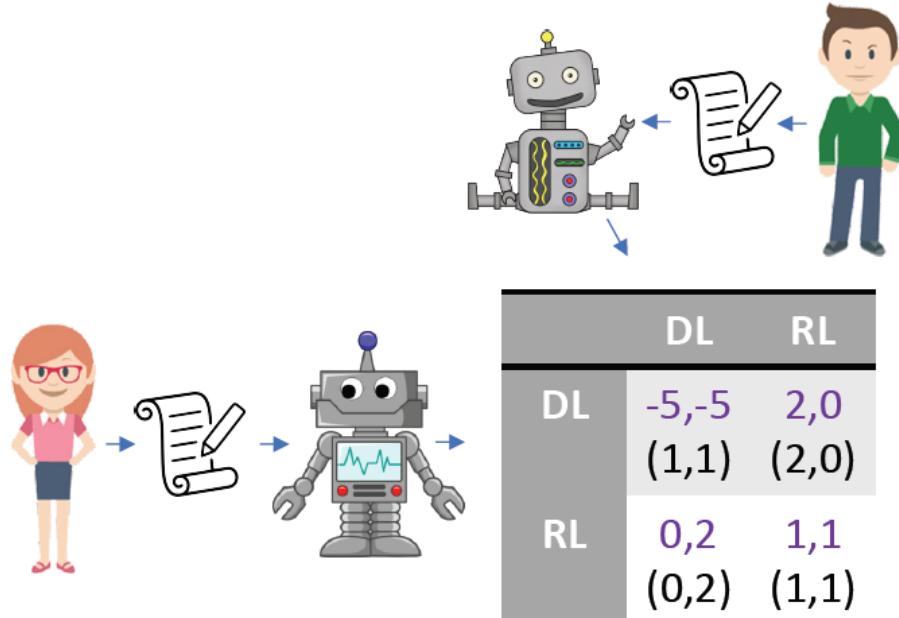
Casper Oesterheld

## Delegated game playing



	DM	RM	DL	RL
DM	-5,-5	2,0	5,-5	5,-5
RM	0,2	1,1	5,-5	5,-5
DL	-5,5	-5,5	1,1	2,0
RL	-5,5	-5,5	0,2	1,1

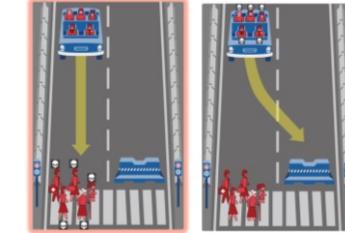
- Representatives are competent at playing games and the original players trust the representatives.  
=> **Default: aligned delegation**
- DL,RL are strictly dominated and therefore never played
- **Equilibrium selection problem**  
=> Pareto-suboptimal outcome (DM,DM) might occur



- Each player's contract says: Play this alternative game if the other player adopts an analogous contract.
- The games are essentially isomorphic.
  - DM ~ DL
  - RM ~ RL
- *Safe Pareto improvement* on the original game: outcome of new game is better for both players with certainty.

# Conclusion

- AI has traditionally strived for the *homo economicus* model
  - Not just “rational” but also: not distributed, full memory, tastes exogenously determined
- Not always appropriate for AI!
- Need to think about choosing objective function
- ... with strategic ramifications in mind
- May not retain / share information across all nodes
- → new questions about how to form beliefs and make decisions
- Social choice, decision, and game theory provide solid foundation to address these questions



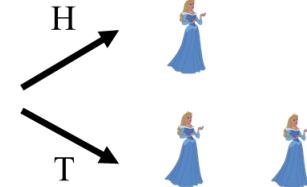
1, 1	-2, 3
3, -2	0, 0

1, 1	-2, 3
3, -2	0, 0

1, 1	-2, 3
3, -2	0, 0



Sunday Monday Tuesday



**THANK YOU FOR YOUR ATTENTION!**

