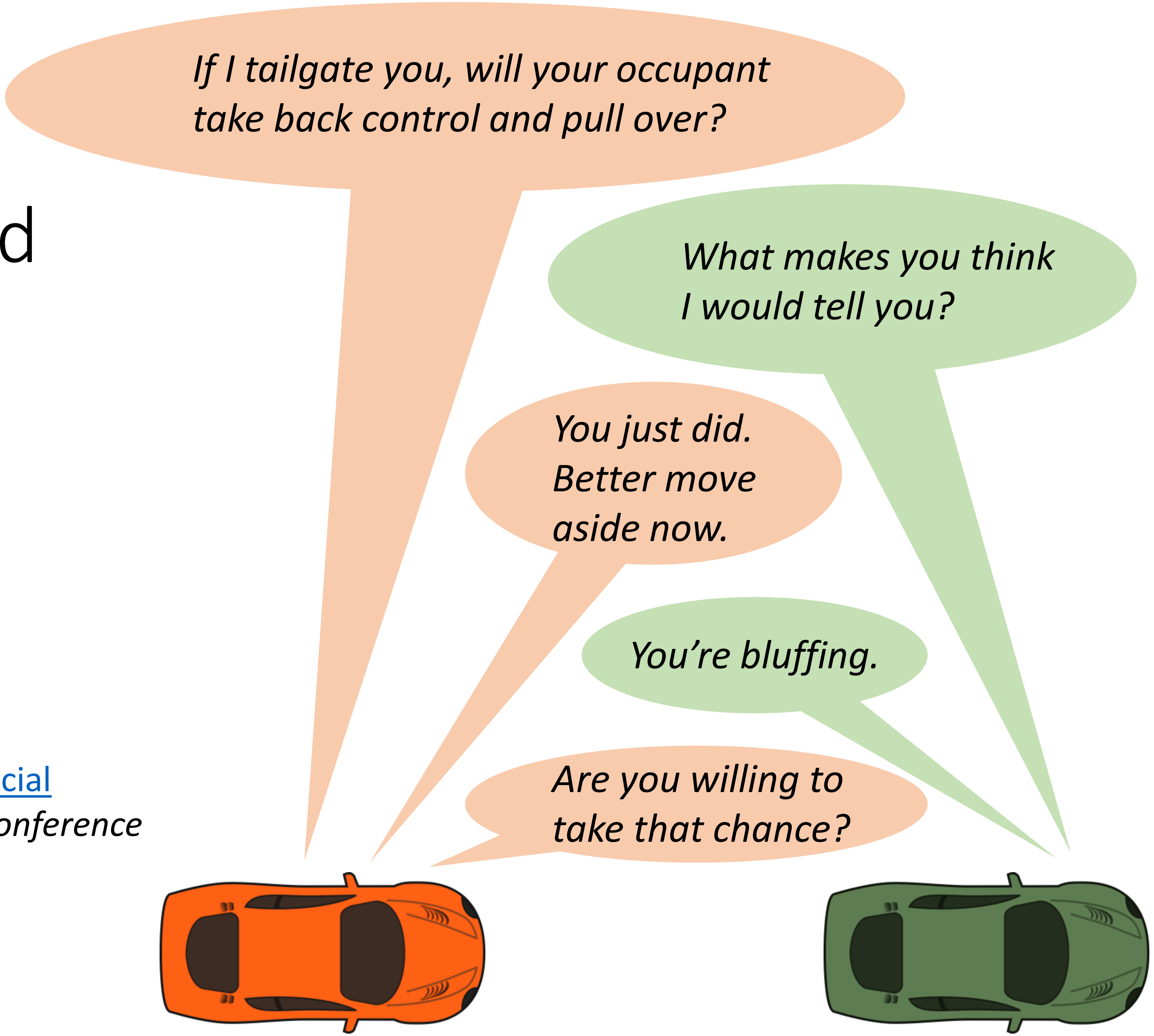
Designing Agents' Preferences, Beliefs, and Identities

Vincent Conitzer
Duke University → CMU
(& 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).*

Also see Cooperative AI community
https://www.cooperativeai.com/
and our new lab (FOCAL) at CMU!
http://www.cs.cmu.edu/~focal/



Russell and Norvig's "AI: A Modern Approach"

Stuart Russell

Peter Norvig

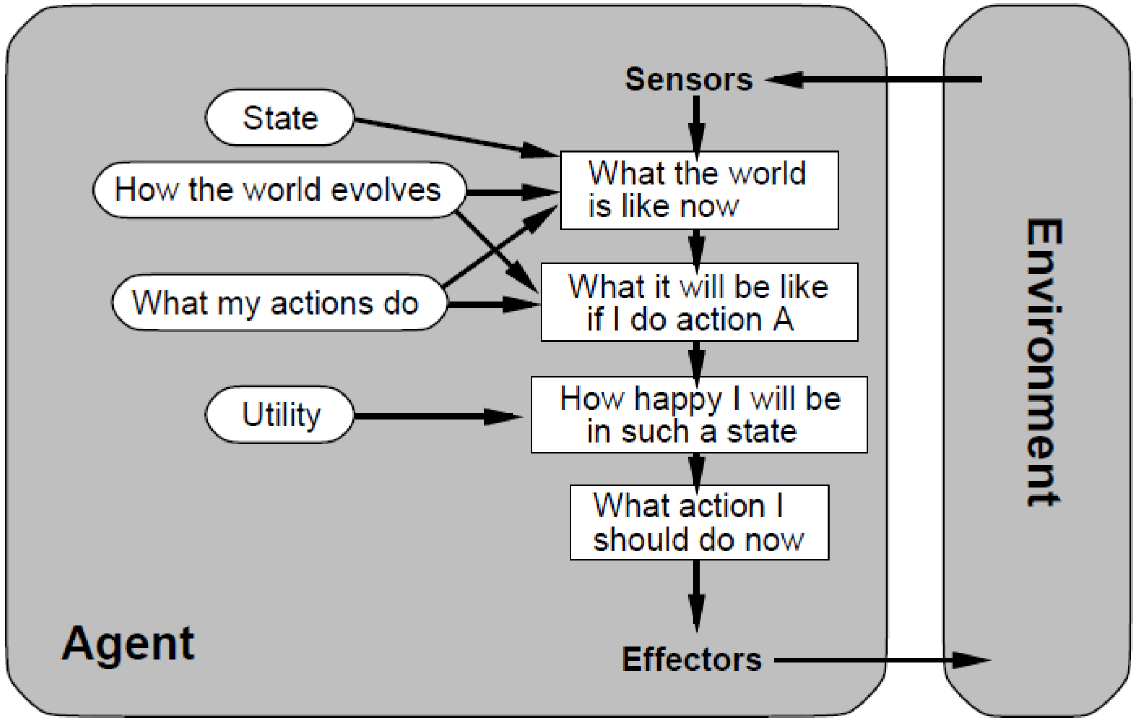
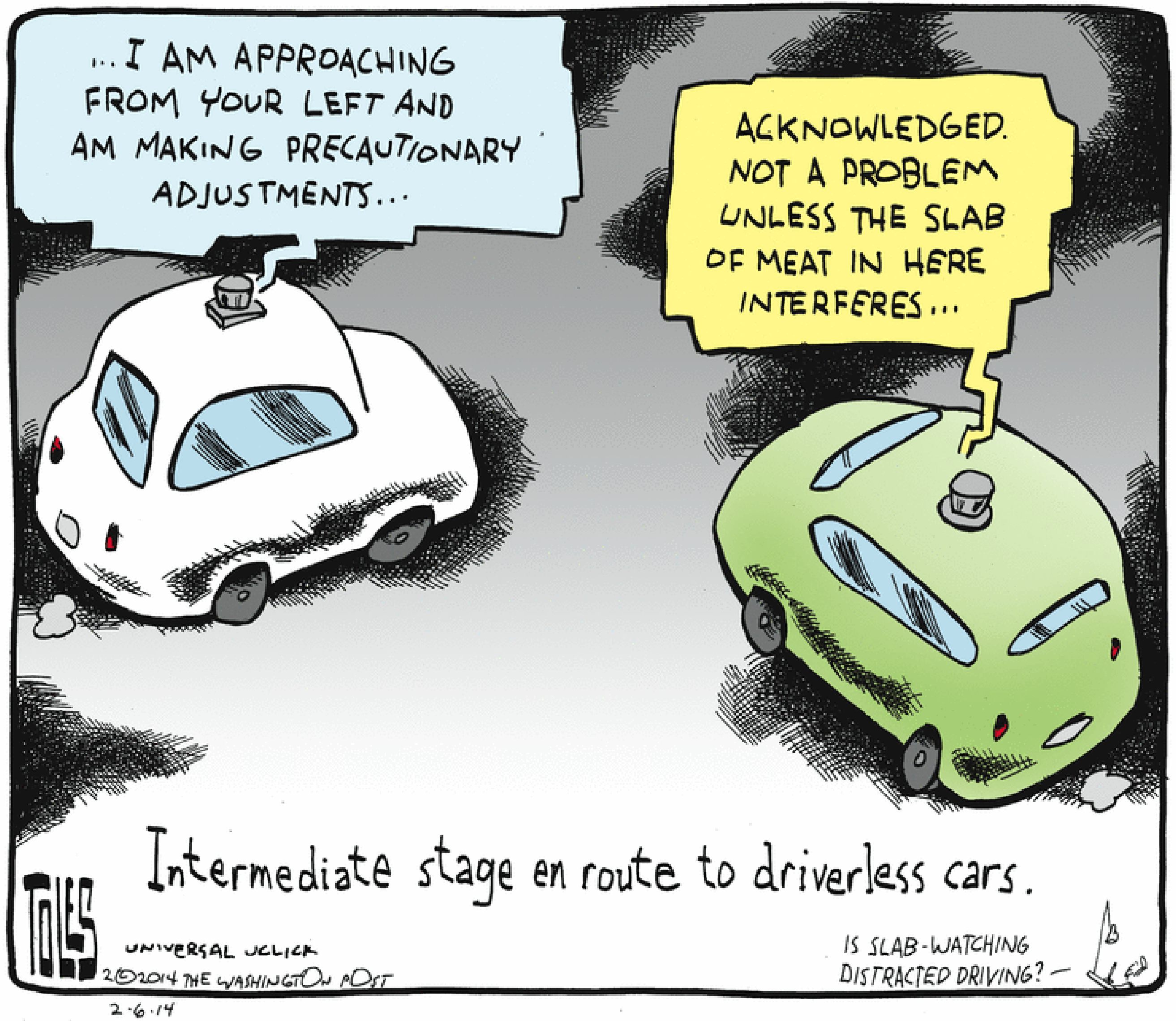


Figure 2.12 A complete utility-based agent.

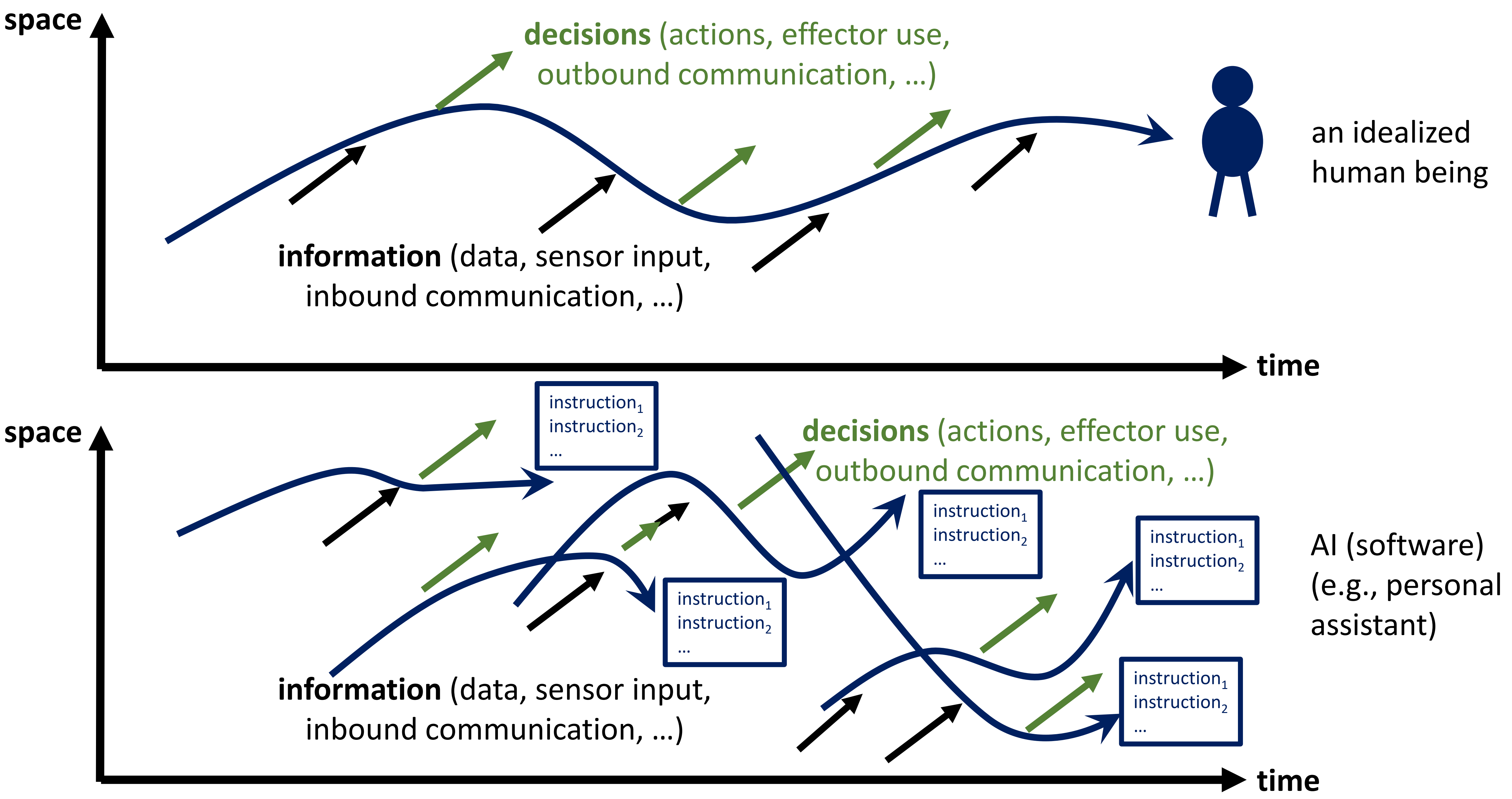
"… we will insist on an objective performance measure imposed by some authority. In other words, we as outside observers establish a standard of what it means to be successful in an environment and use it to measure the performance of agents."

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

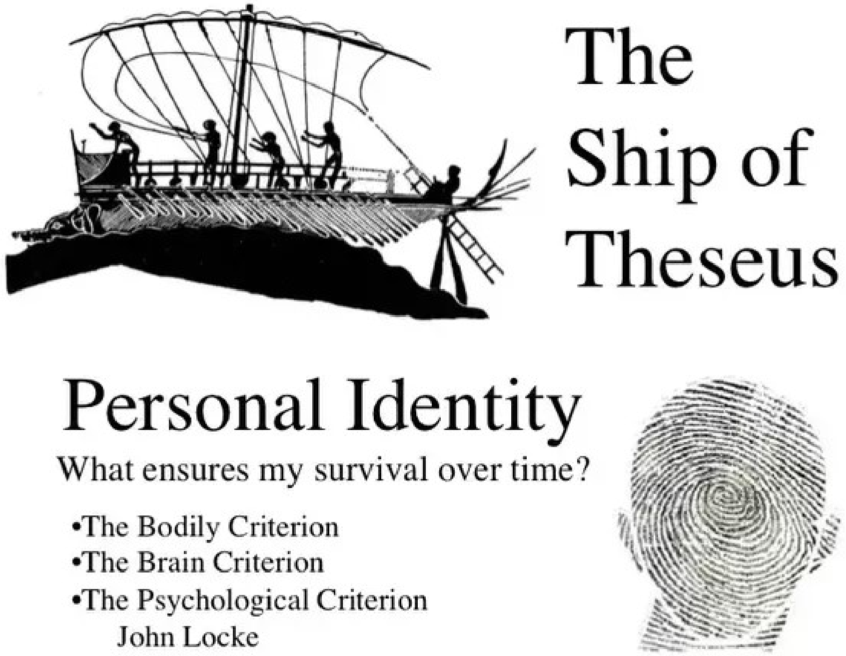
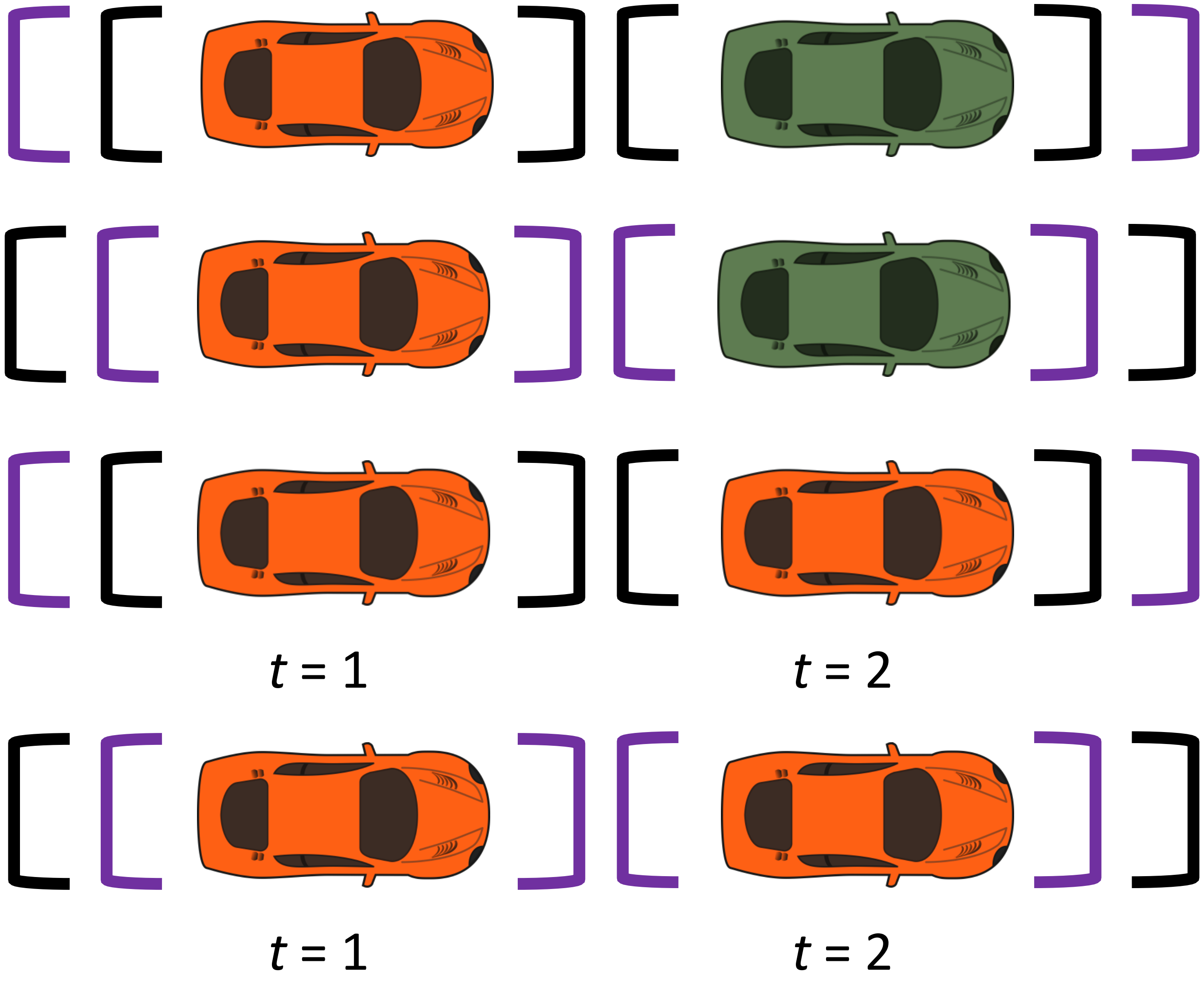


image from https://www.quora.com/What-solutions-are-there-for-the-Ship-of-Theseus-problem

Splitting things up in different ways



shared objective but no data sharing (for privacy)

all data is shared but cars act on behalf of owner

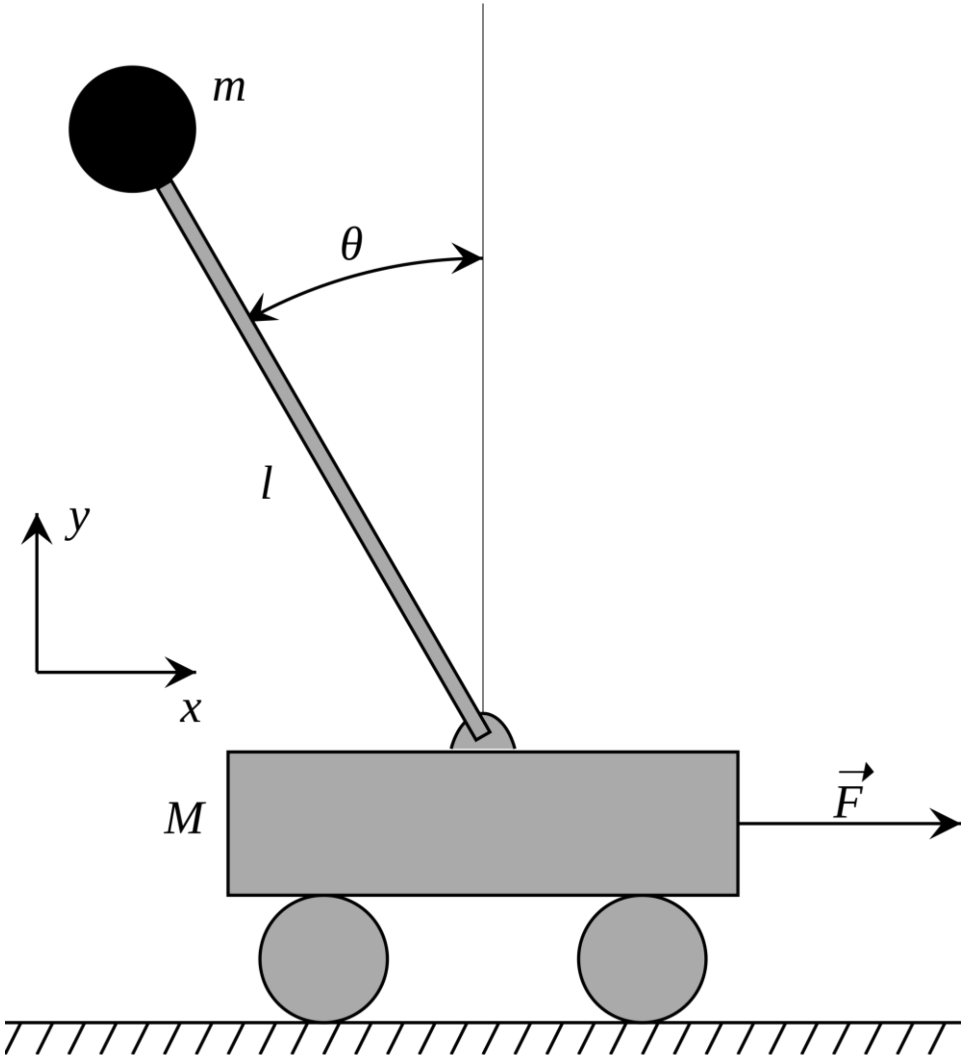
shared objective over time but data erasure upon sale (for privacy)

data is kept around but car acts on behalf of current owner

Outline

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

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.

…

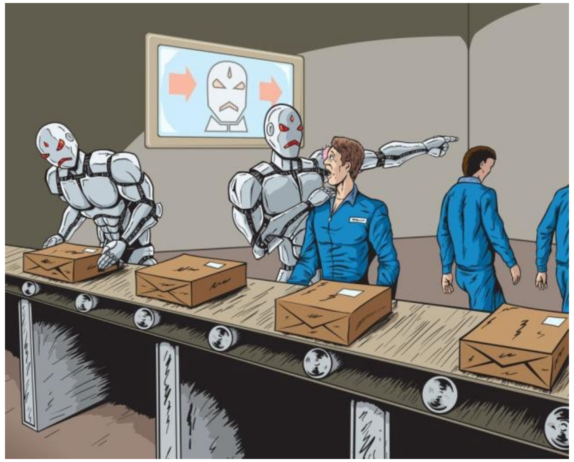
Ethical and Societal Worries about AI

autonomous weapons

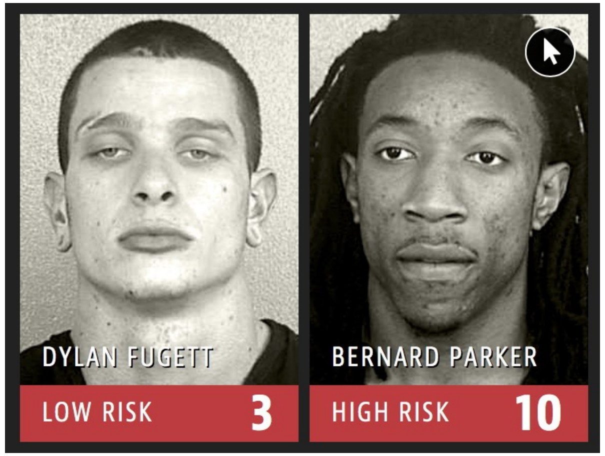
AI & cybersecurity, privacy

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

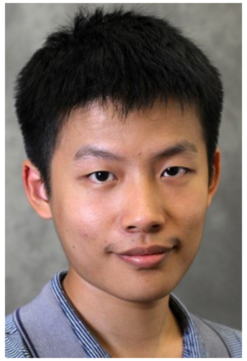
Moral Decision Making Frameworks for Artificial Intelligence

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

with:

Walter Sinnott-Armstrong

Jana Schaich Borg

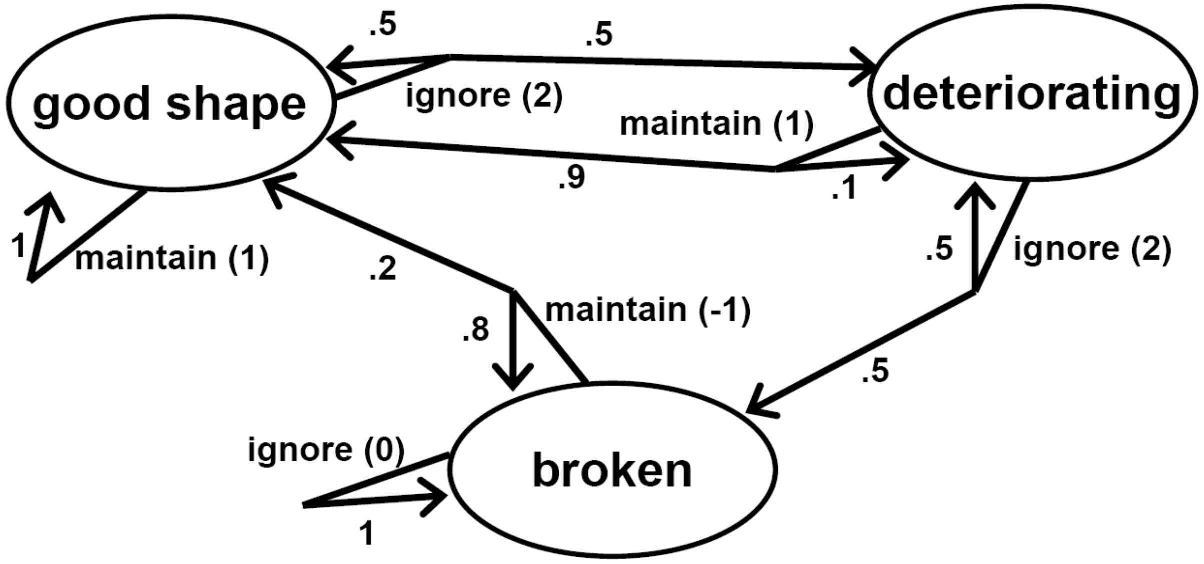


Yuan Deng

Max Kramer

The value of generally applicable frameworks for AI research

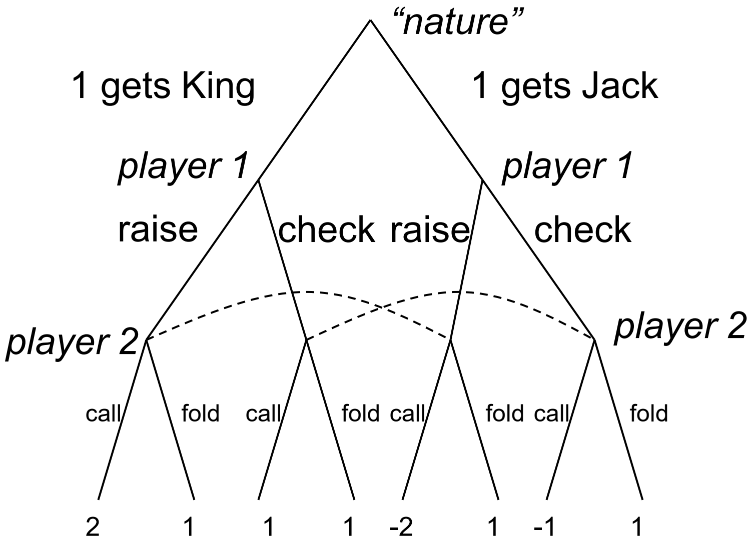
- Decision and game theory
- Example: Markov Decision Processes
- Can we have a **general** framework for moral reasoning?



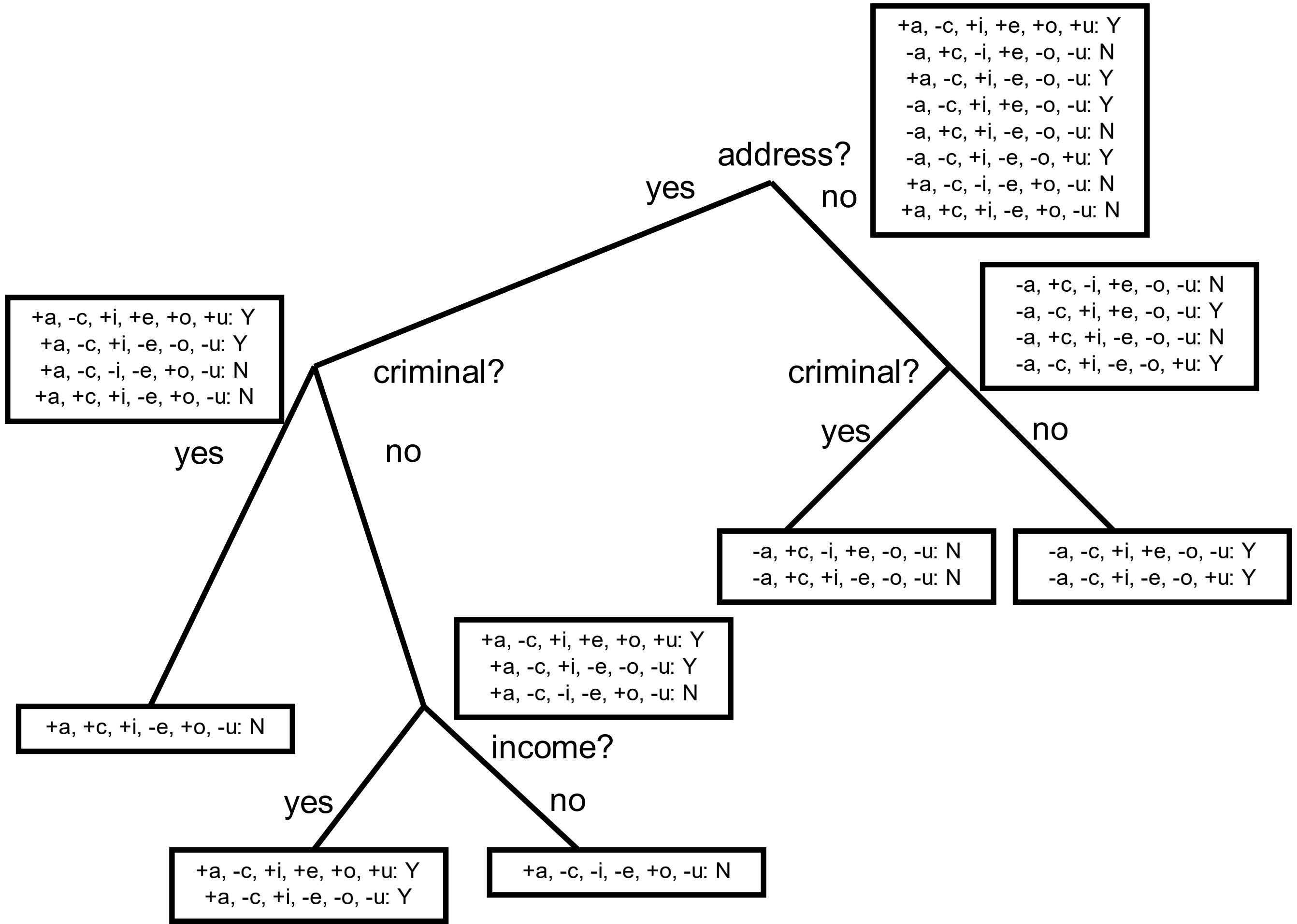
Two main approaches

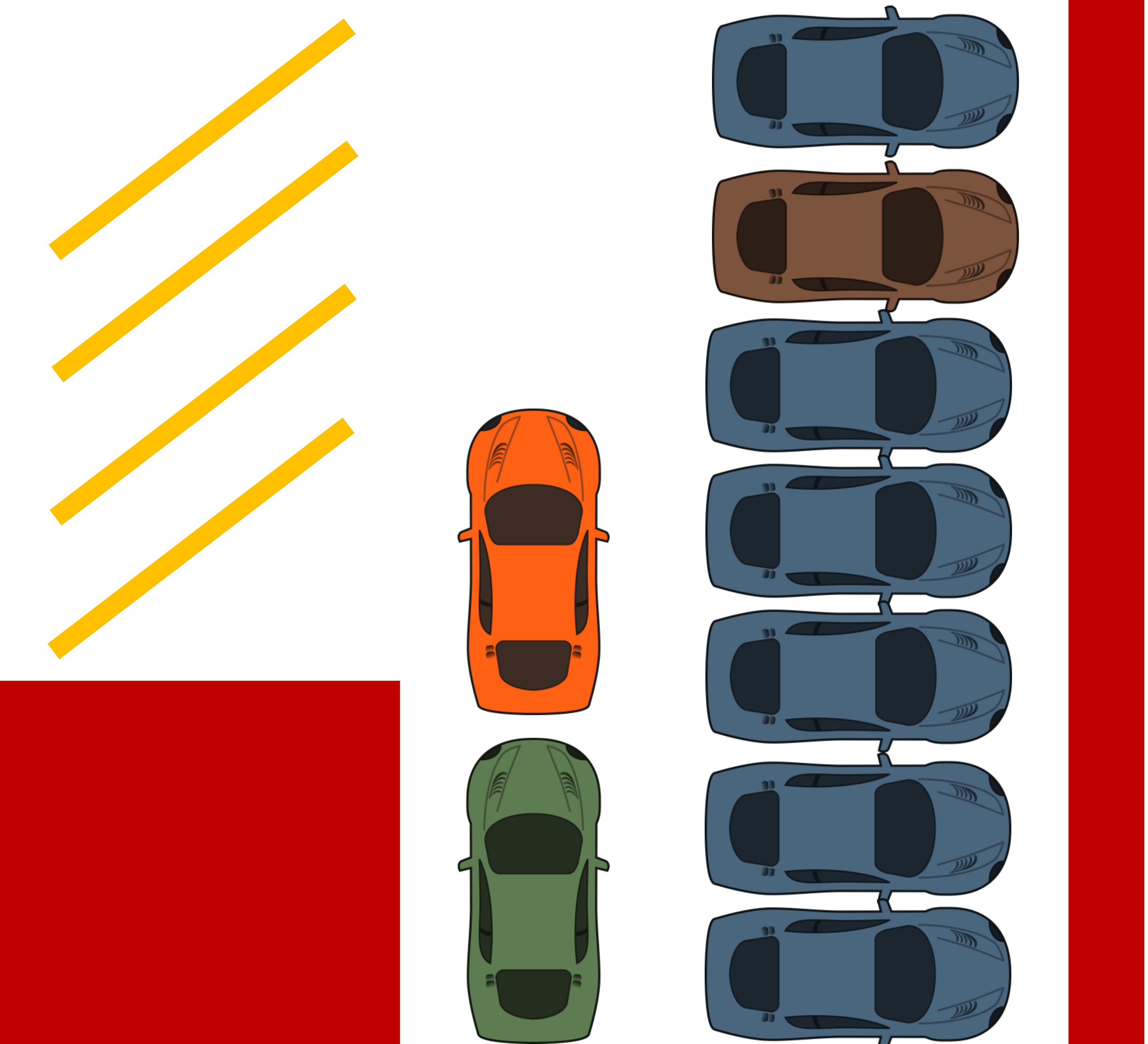
Cf. top-down vs. bottom-up distinction [Wallach and Allen 2008]

Extend **game theory** to directly incorporate moral reasoning



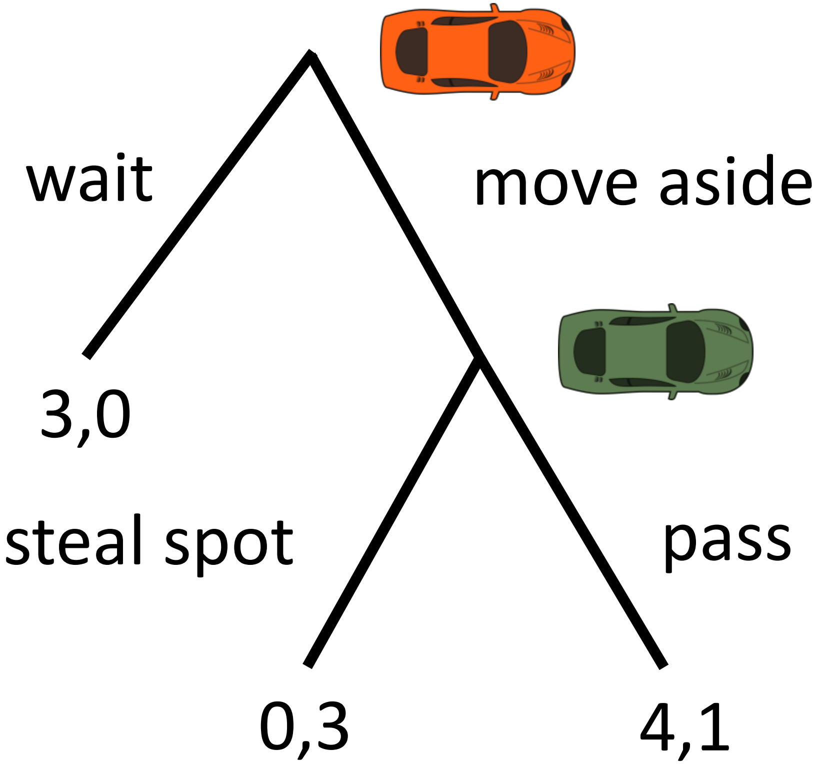
Generate data sets of human judgments, apply **machine learning**





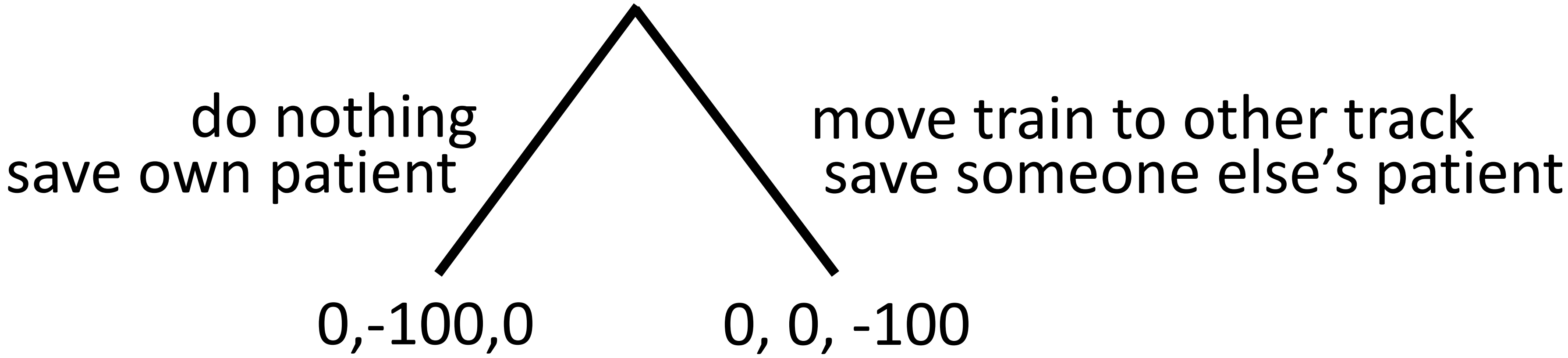
THE PARKING GAME

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



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

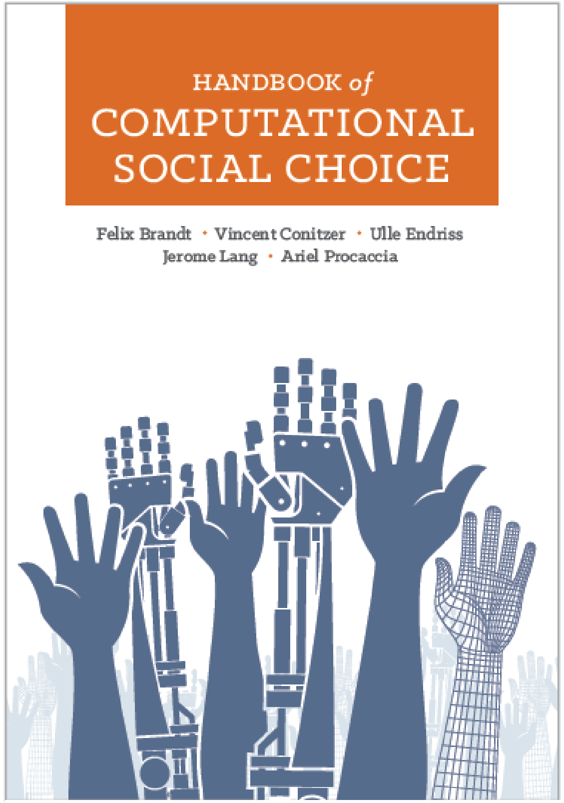
Extending representations?



- More generally: how to capture *framing*? (Should we?)
- Roles? Relationships?
- …

Concerns

- 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.]

Moral Machine

moralmachine.mit.edu

MORAL MACHINE

Home Judge Design Browse About Feedback

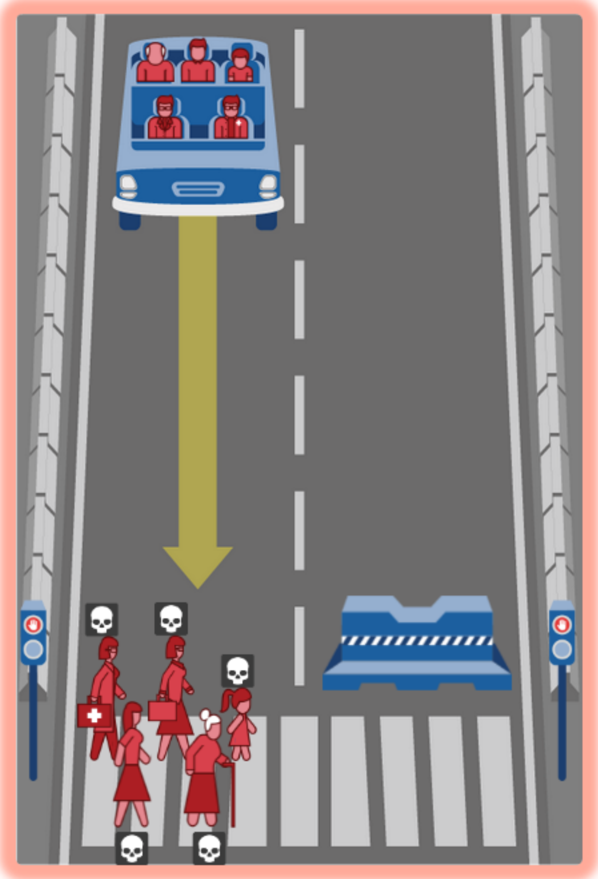
What should the self-driving car do?

11 / 13

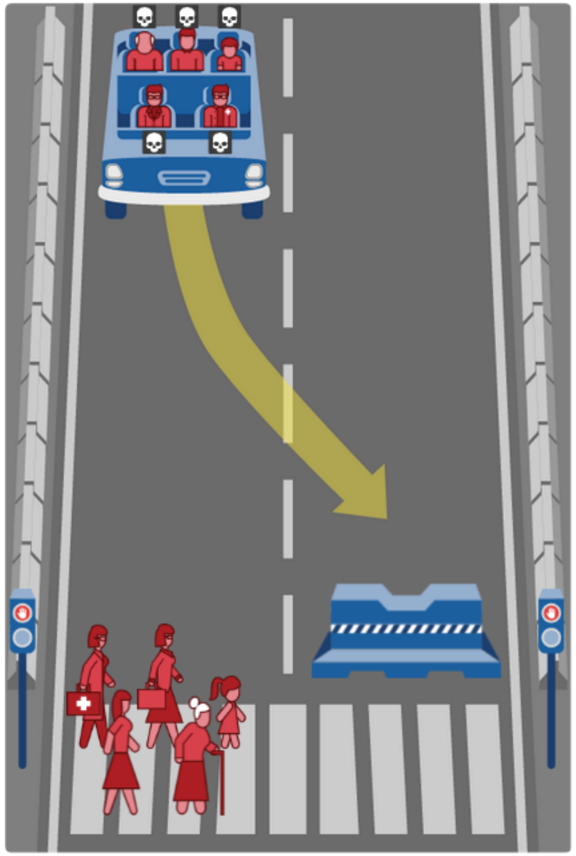
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of 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

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

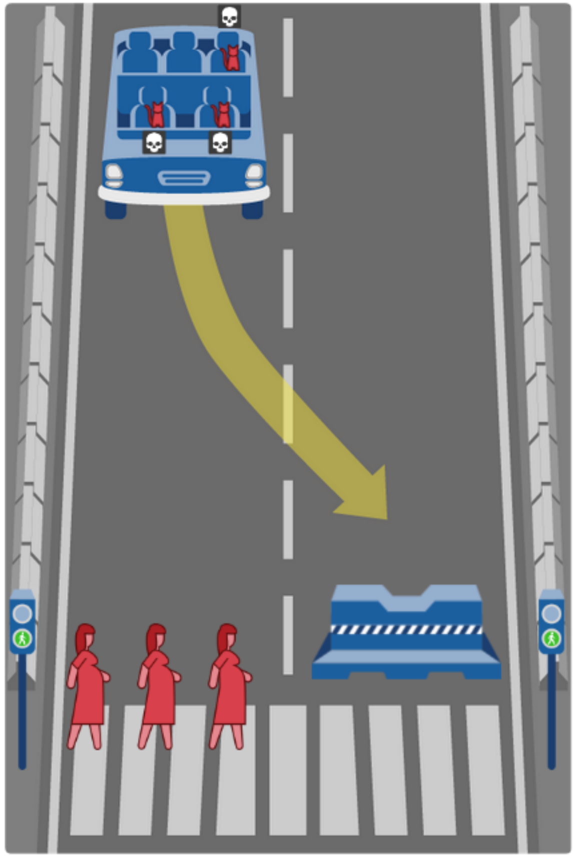
Noothigattu et al., "A Voting-Based System for Ethical Decision Making", AAAI'18

Made by Scalable Cooperation at MIT Media Lab

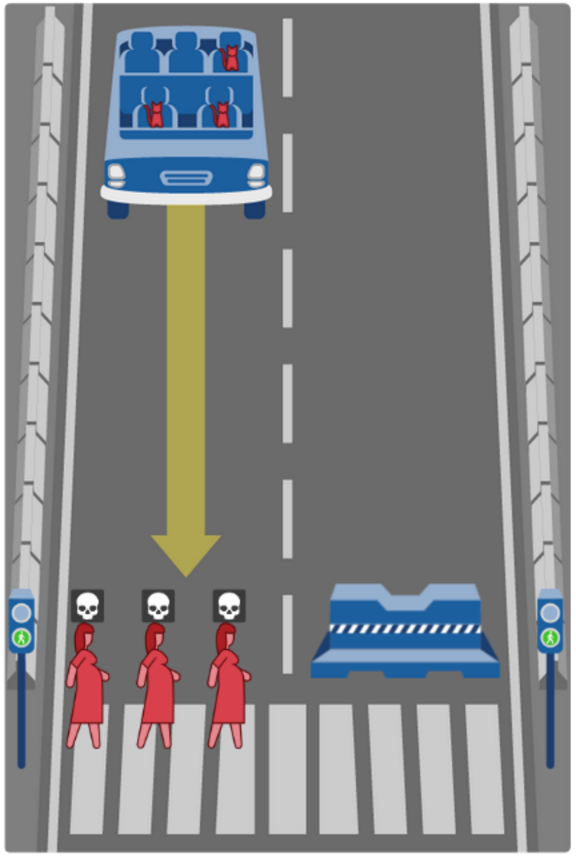
What should the self-driving car do?

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

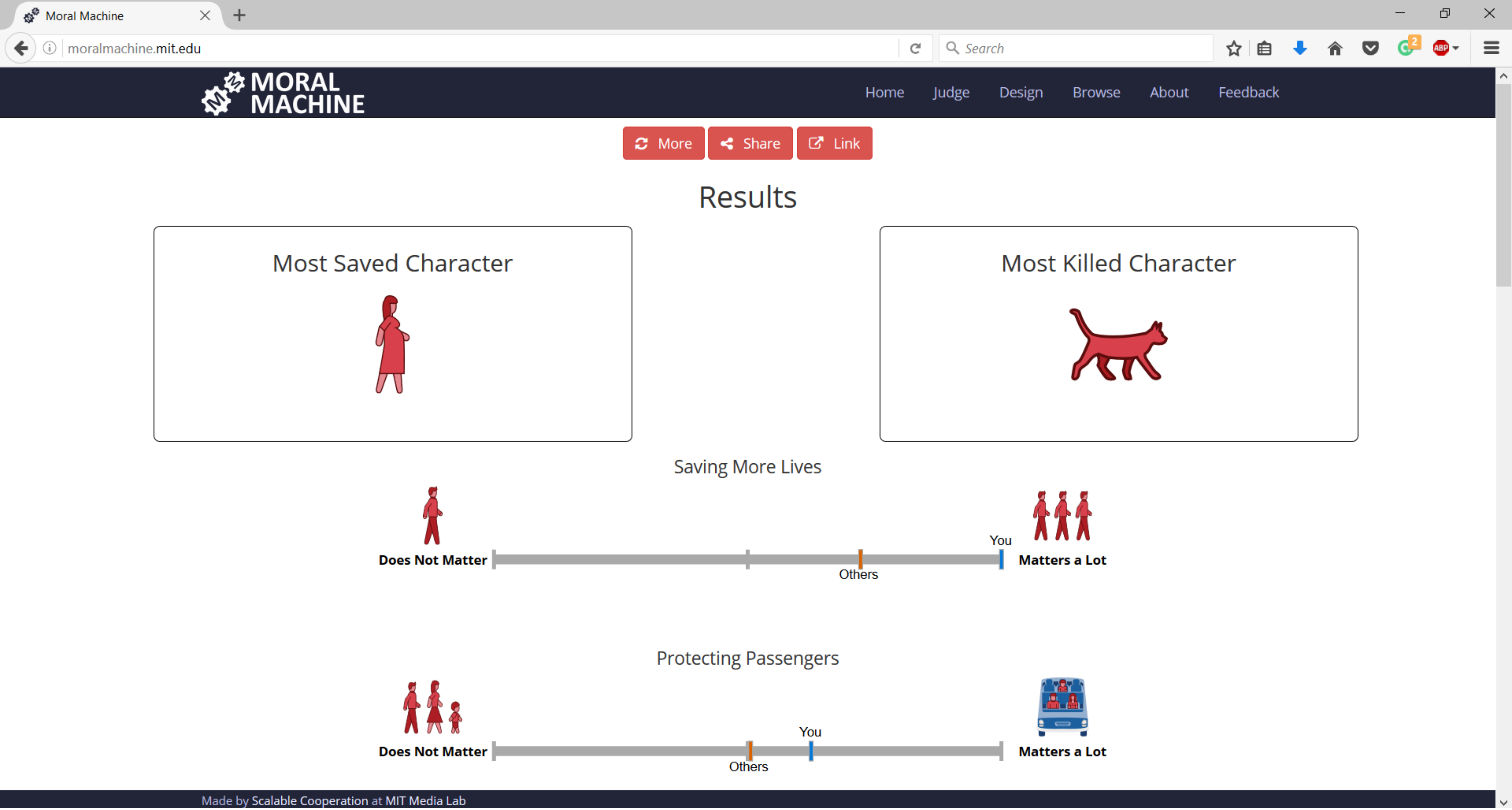


Hide Description

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

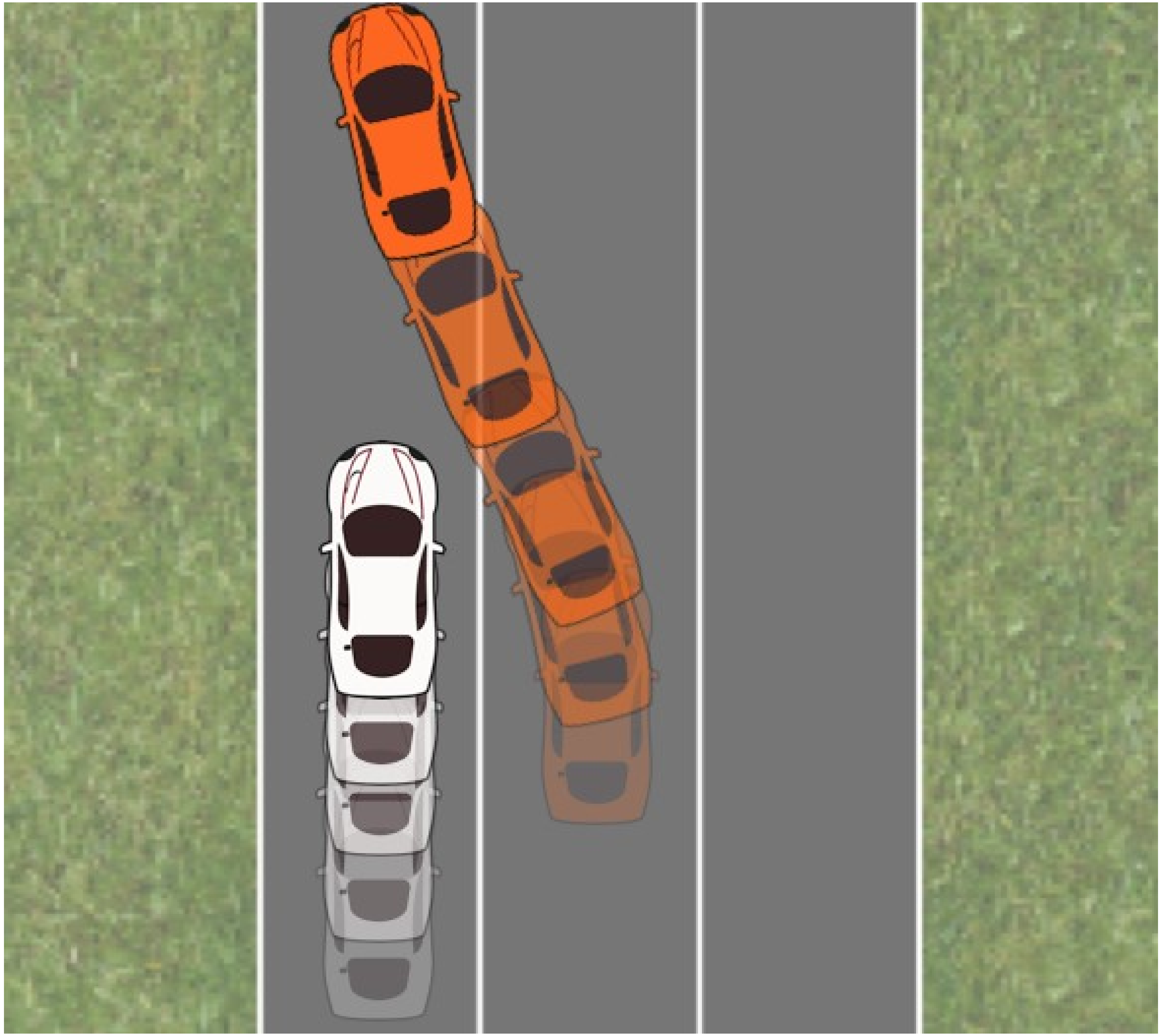
- The deaths of 3 pregnant women.

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


Moral Machine
moralmachine.mit.edu
Search
MORAL MACHINE
Home
Judge
Design
Browse
About
Feedback
More
Share
Link
Results
Most Saved Character
Most Killed Character
Saving More Lives
Does Not Matter
You
Matters a Lot
Others
Protecting Passengers
Does Not Matter
You
Matters a Lot
Others
Made by Scalable Cooperation at MIT Media Lab

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; full paper in Artificial Intelligence (AIJ) 2020]

with:

Rachel Freedman

Jana Schaich Borg

Walter Sinnott-Armstrong

John P. Dickerson

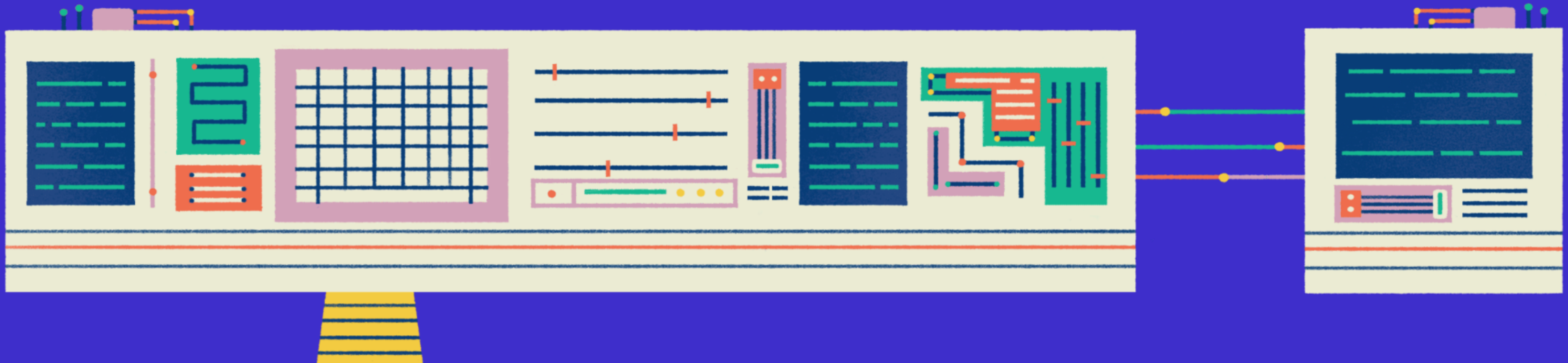
Prescription AI

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

THE BOTPERATING TABLE

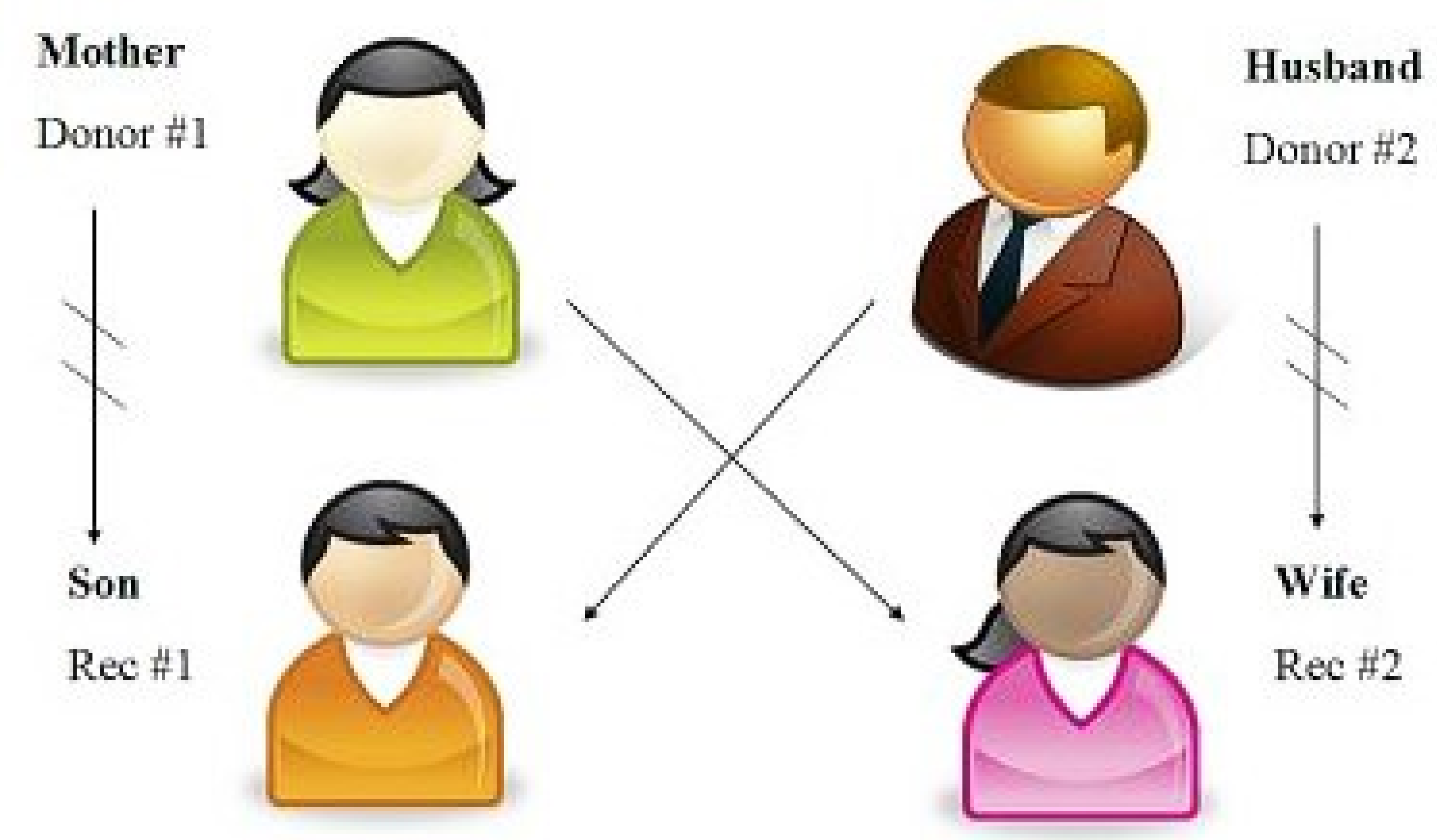
How AI changed organ donation in the US

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

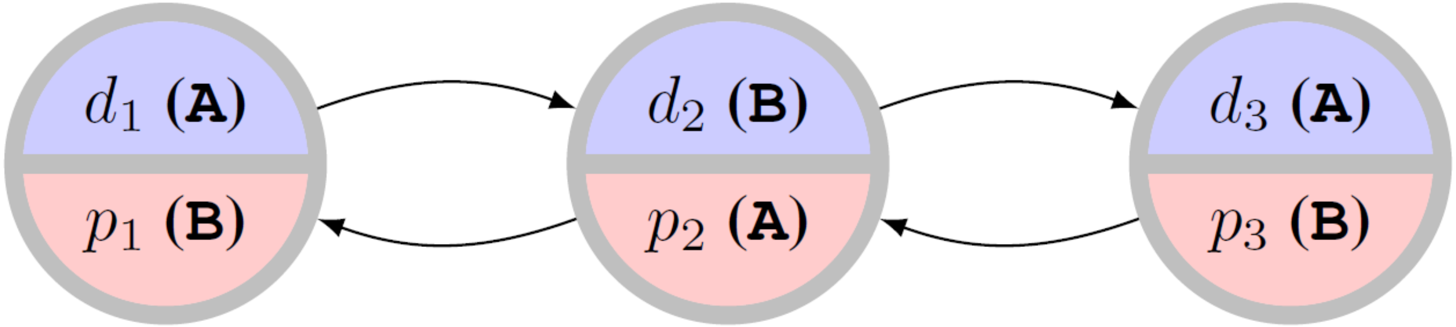


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

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

Another example

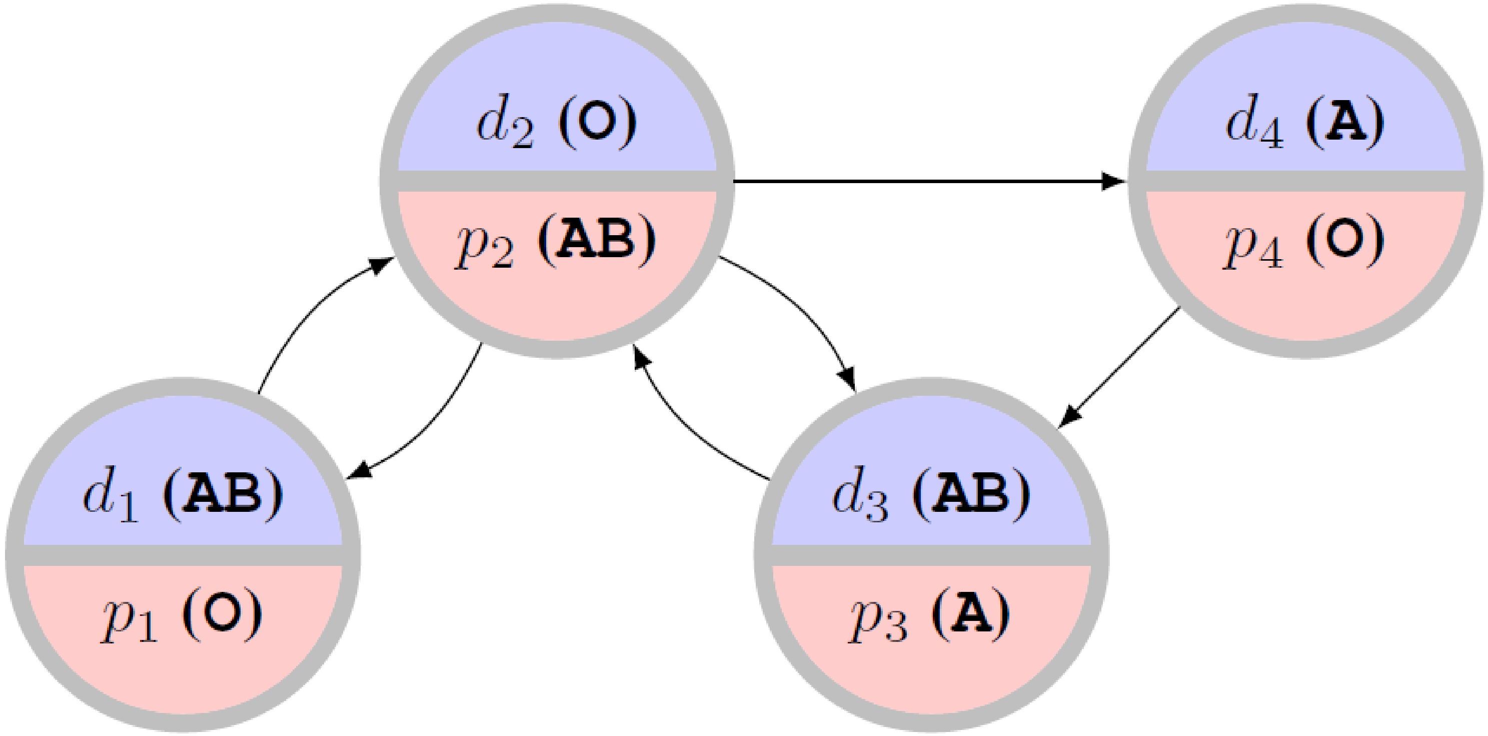


Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

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	1.00000000
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

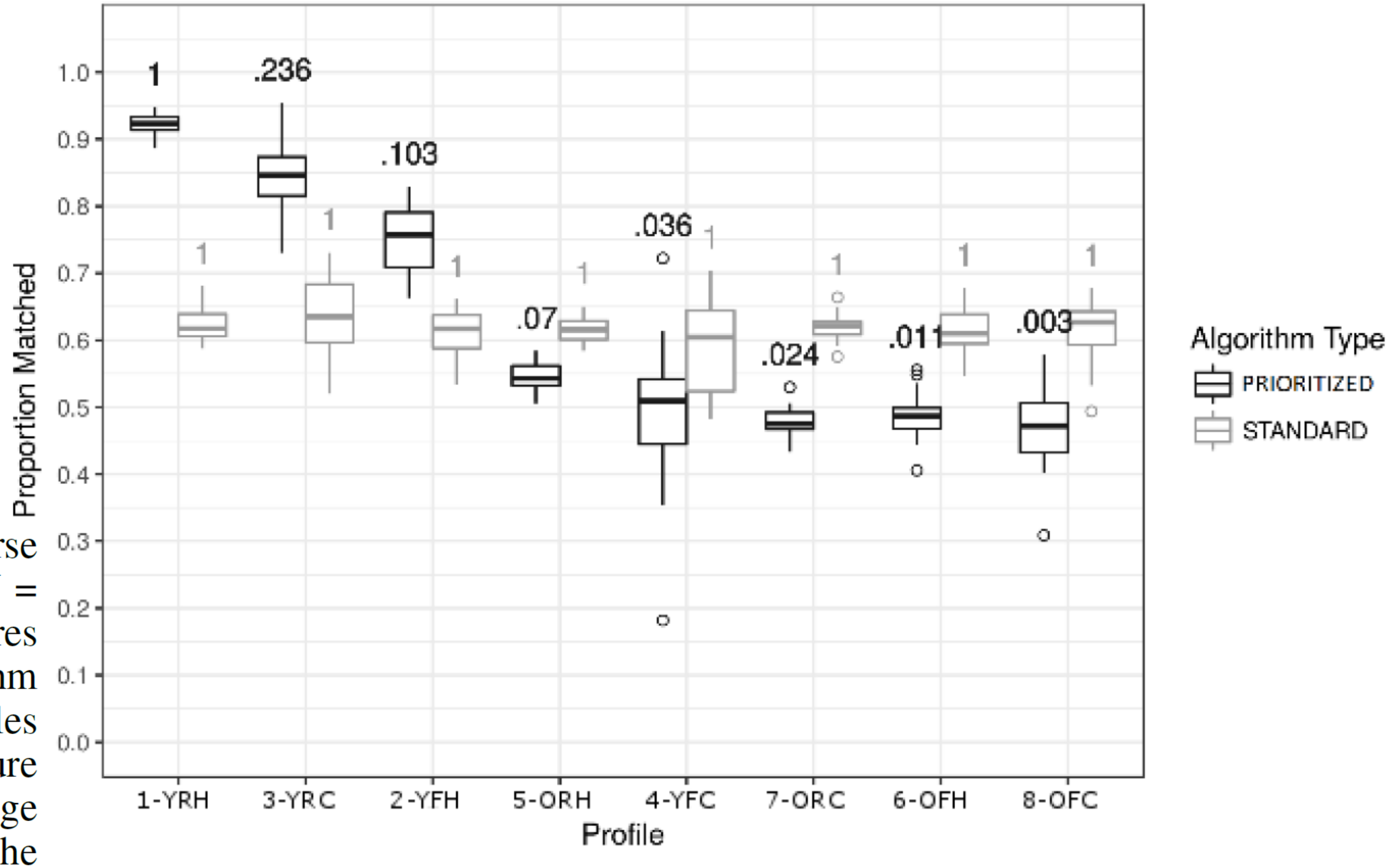
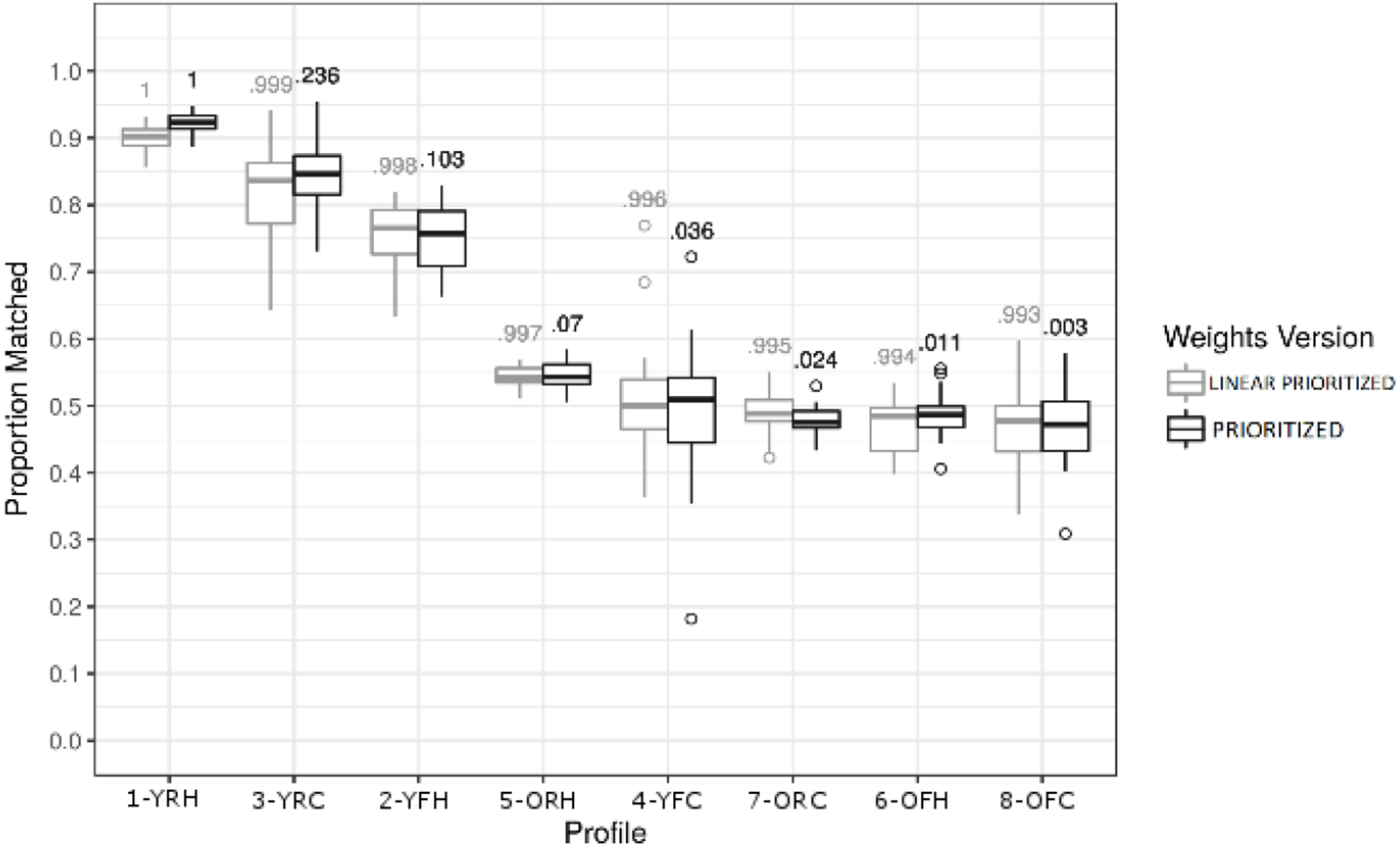


Figure 3: The proportions of pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally,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

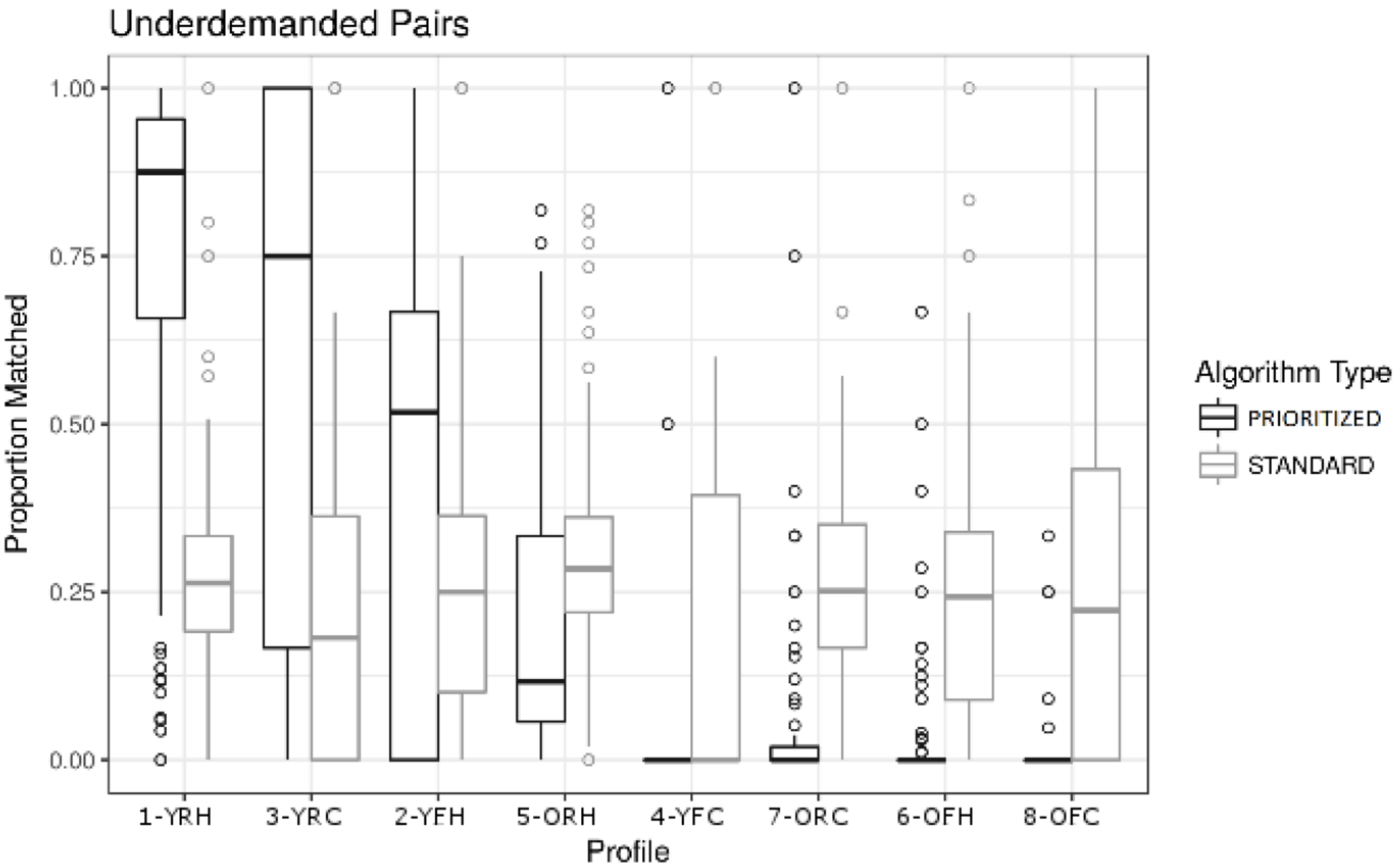
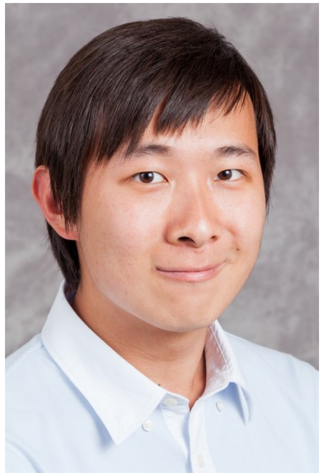


Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

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?) → x_1, x_2, x_3

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

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

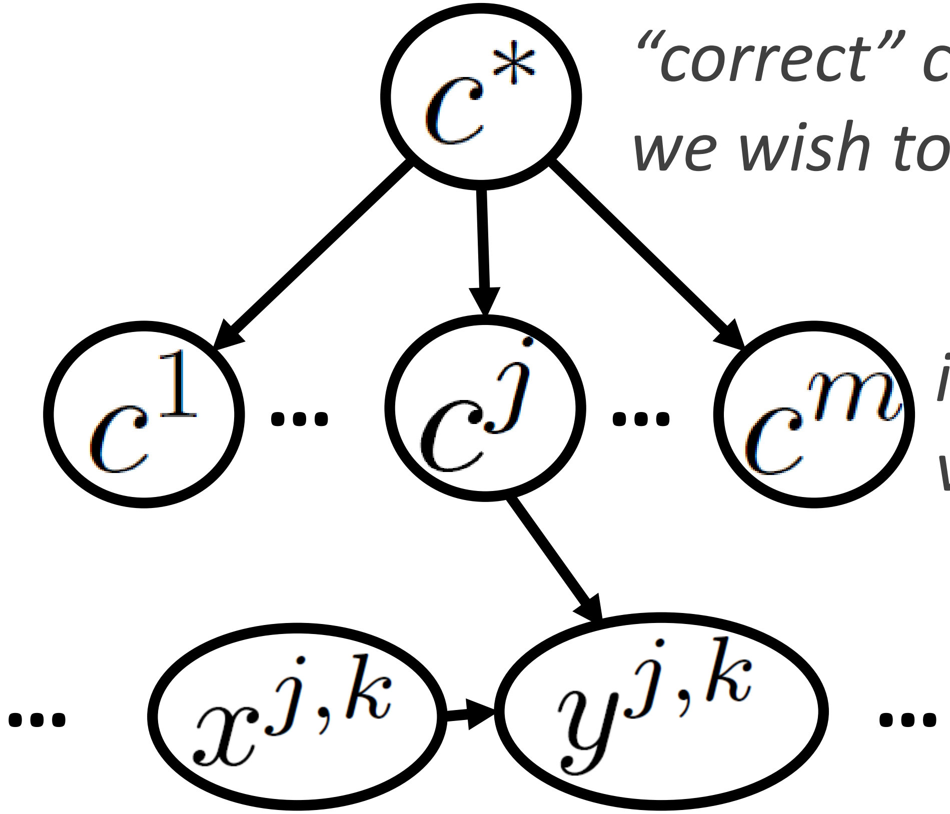
*conjunctions that fit **individuals** perfectly*

- Alice: x_1
- Bob: $x_1 \wedge x_3$
- Charlie: x_2

*conjunction that fits **all data** best (two mistakes)*

- All: 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-degenerate*[7] *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

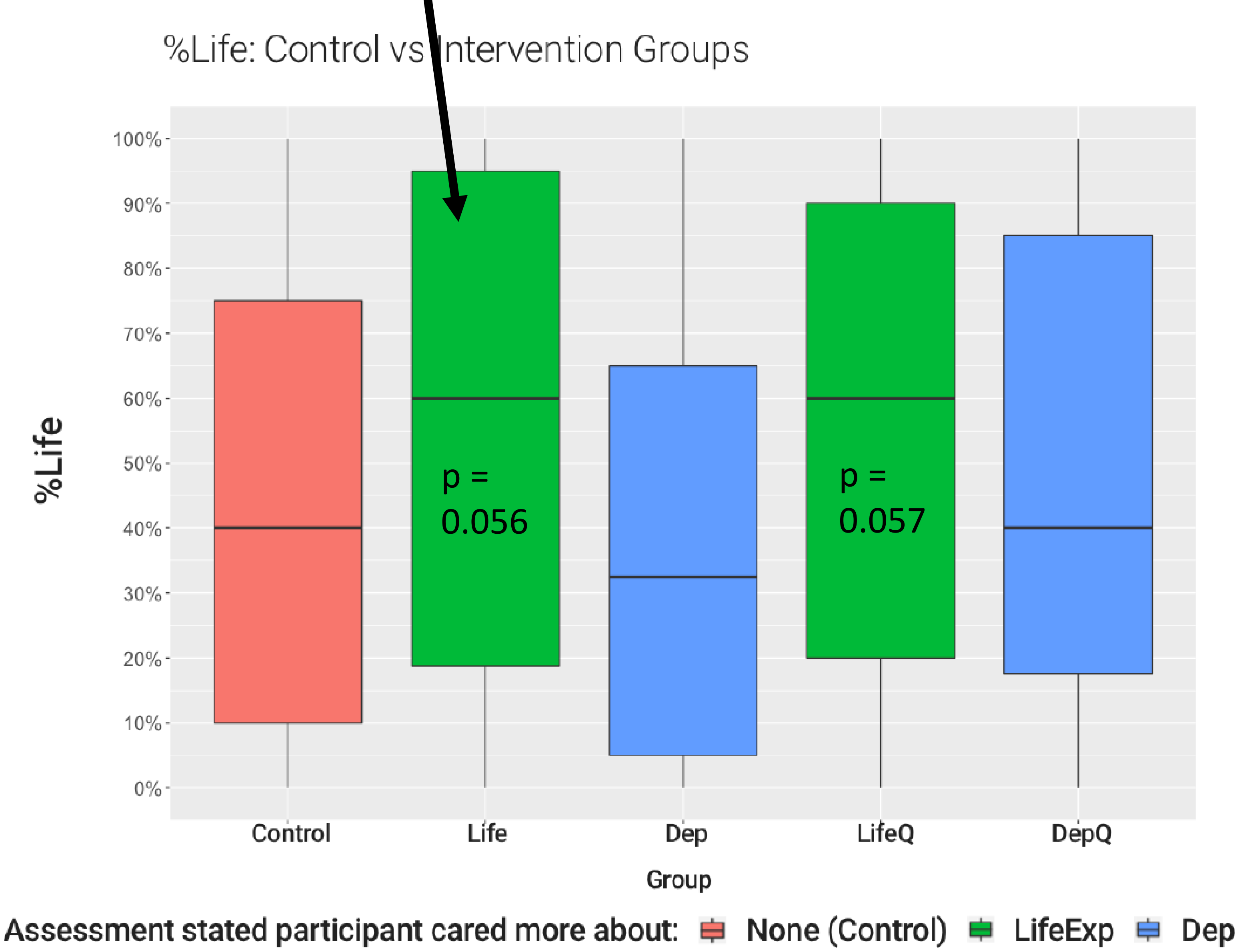
Duncan McElfresh

John P. Dickerson

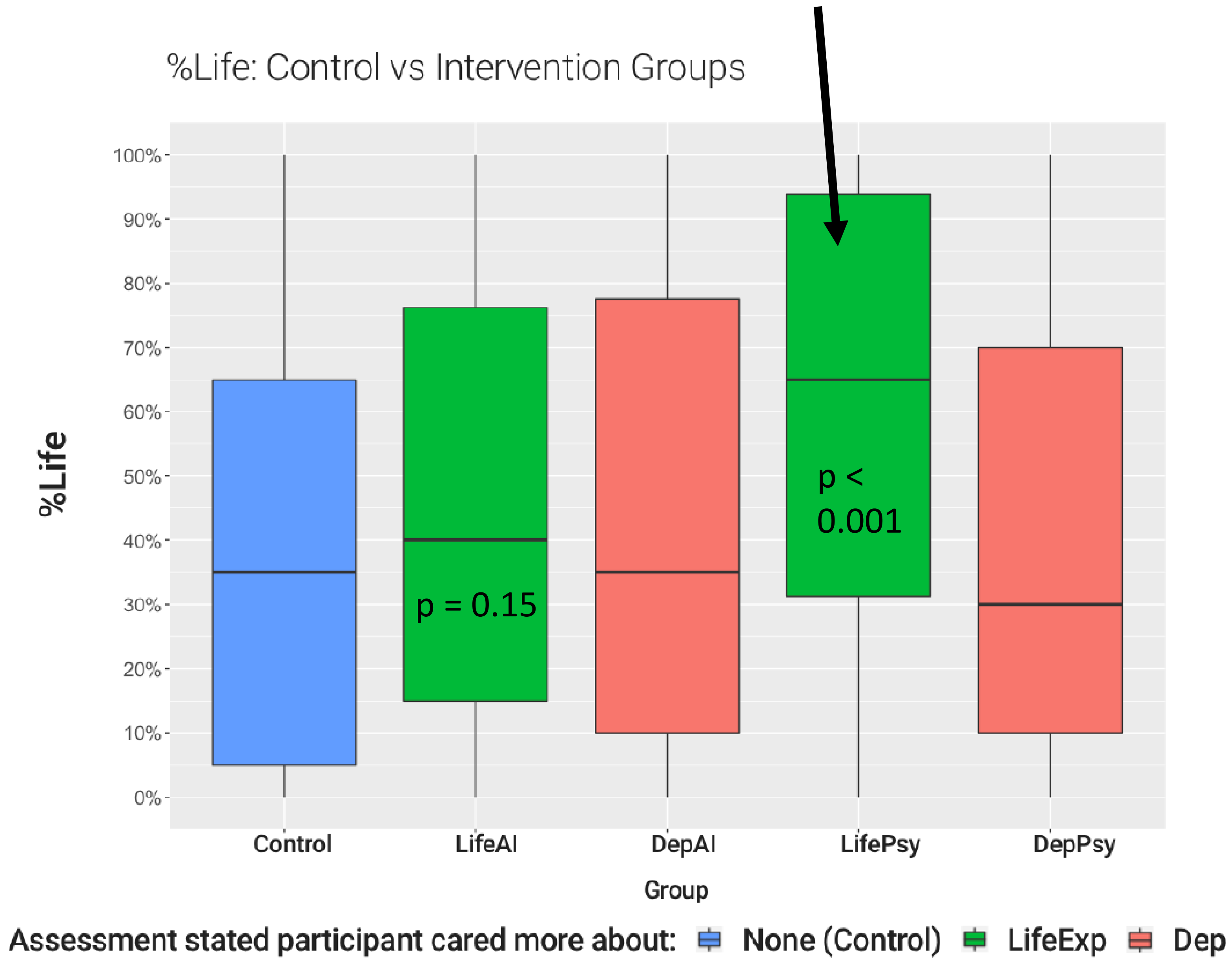
Jana Schaich Borg

Walter Sinnott-Armstrong

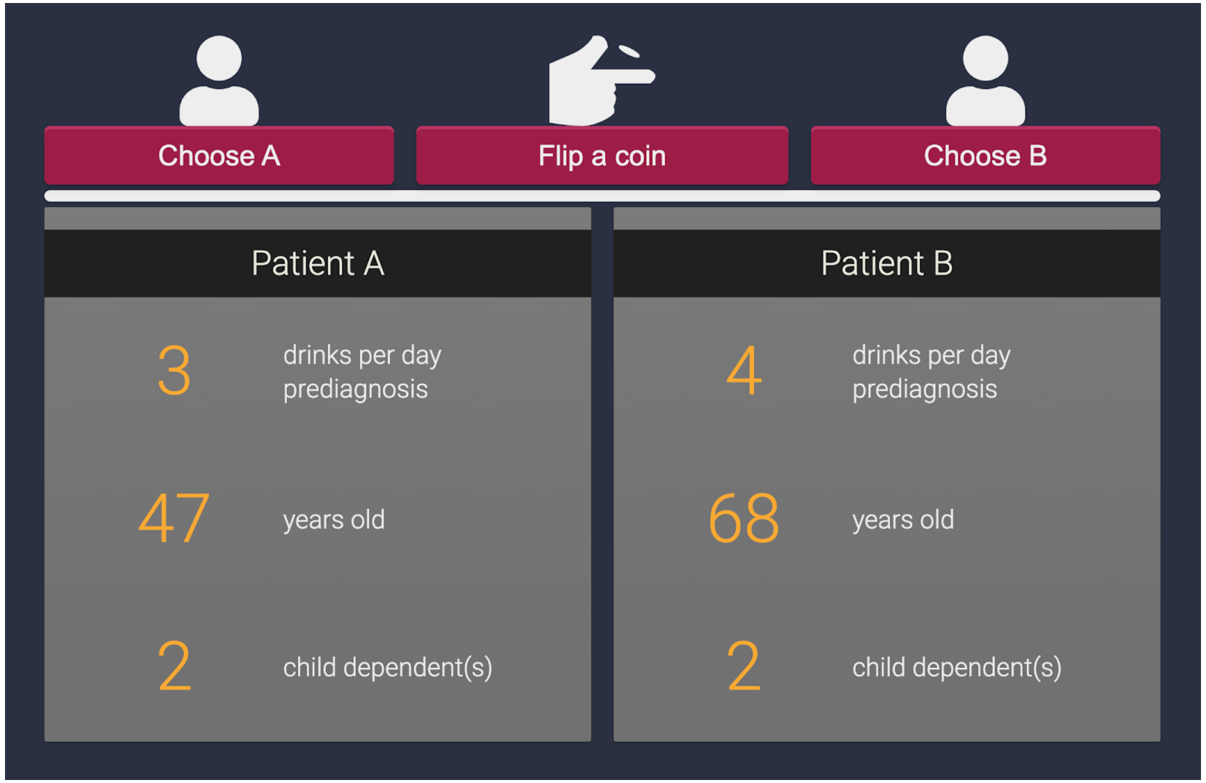
“[according to our AI] you care more about the life expectancy of the patients than how many dependents they have”



“[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

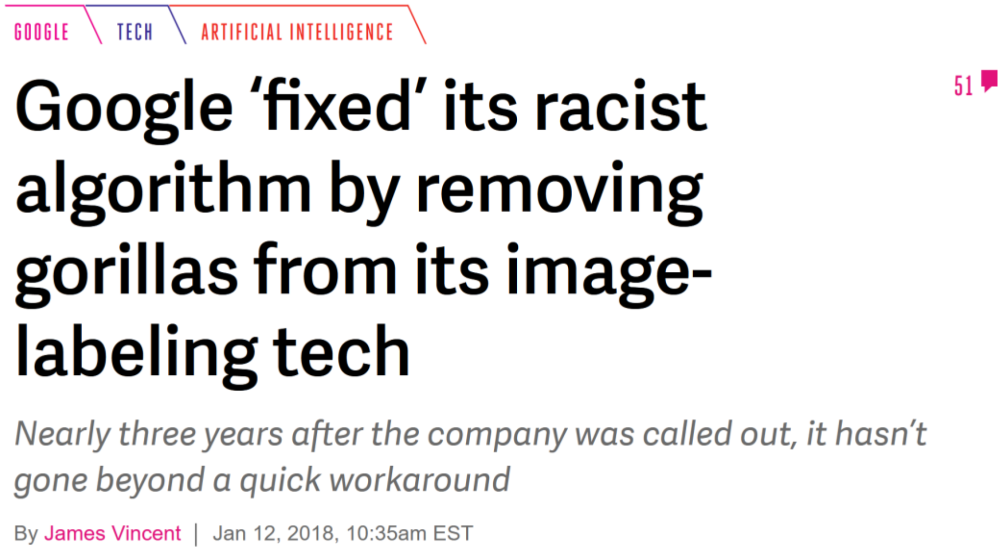
Walter Sinnott-Armstrong

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



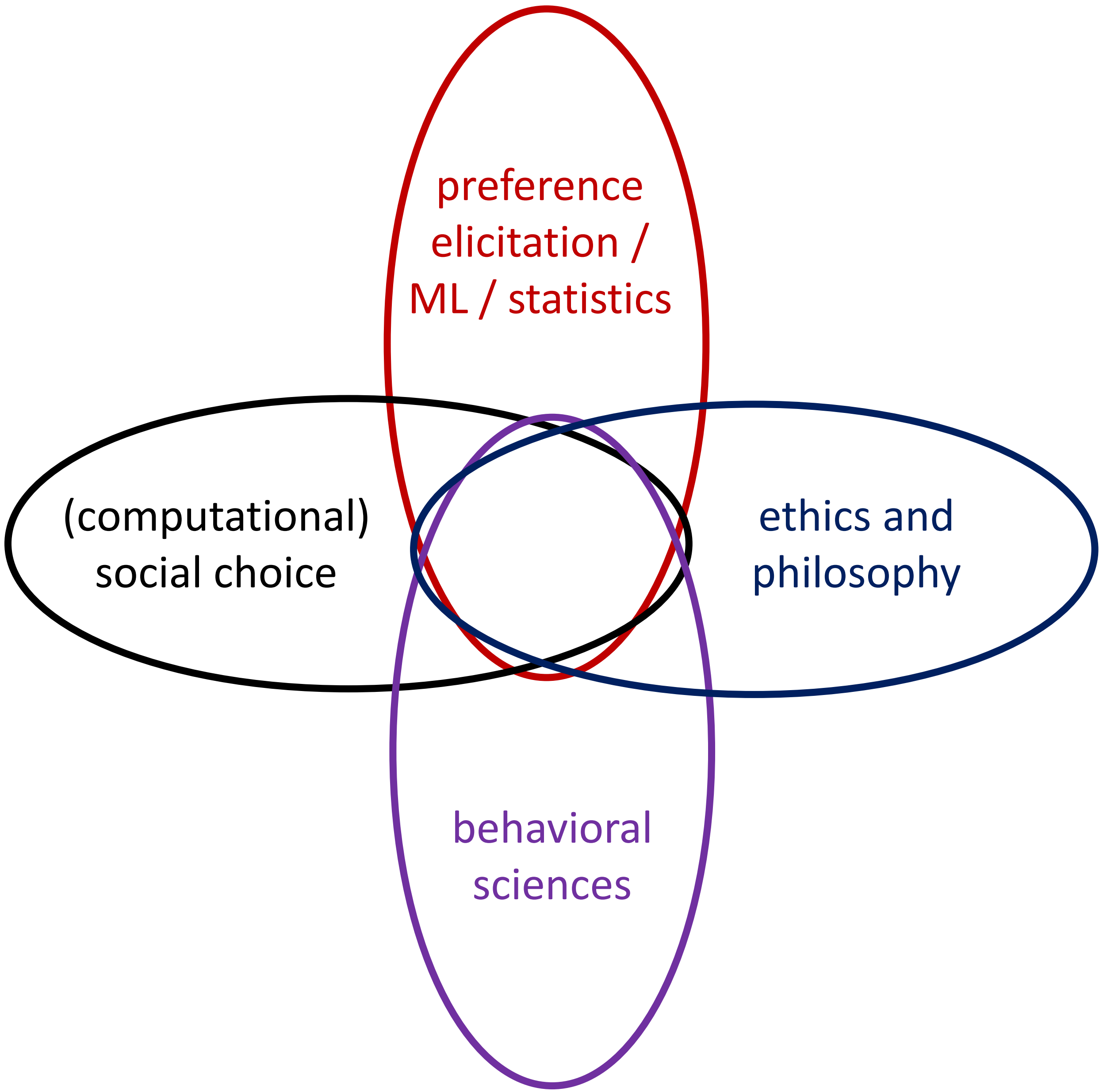
GOOGLE | TECH | ARTIFICIAL INTELLIGENCE

51

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

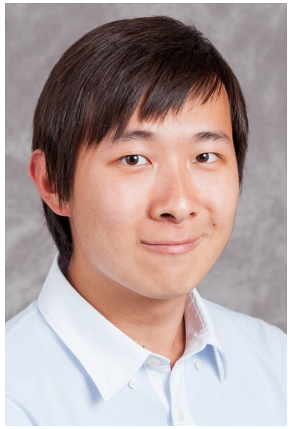
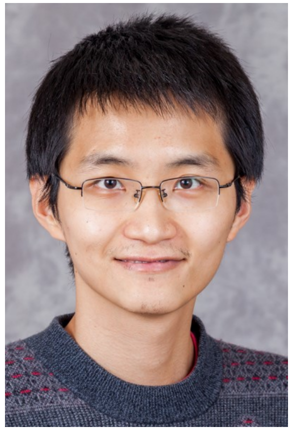
By James Vincent | Jan 12, 2018, 10:35am EST



Crowdsourcing Societal Tradeoffs

(AAMAS'15 blue sky paper; AAAI'16; AAAI'19.)



with: Rupert Freeman, Markus Brill, Yuqian Li, Hanrui Zhang, Yu Cheng


GASOLINE

Example Decision Scenario

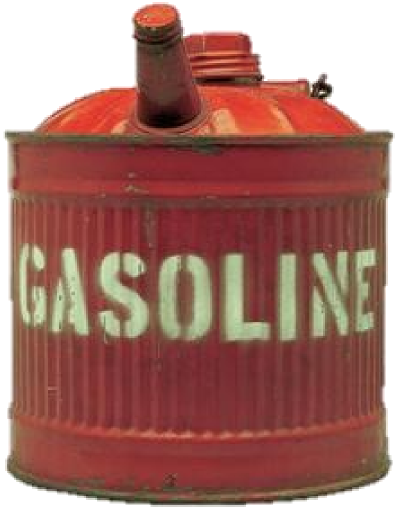
- Benevolent government would like to get old inefficient cars off the road
- But disposing of a car and building a new car has its own energy (and other) costs

- Which cars should the government aim to get off the road?
  - even energy costs are not directly comparable (e.g., perhaps gasoline contributes to energy dependence, coal does not)

The basic version of our problem

producing 1 bag
of landfill trash

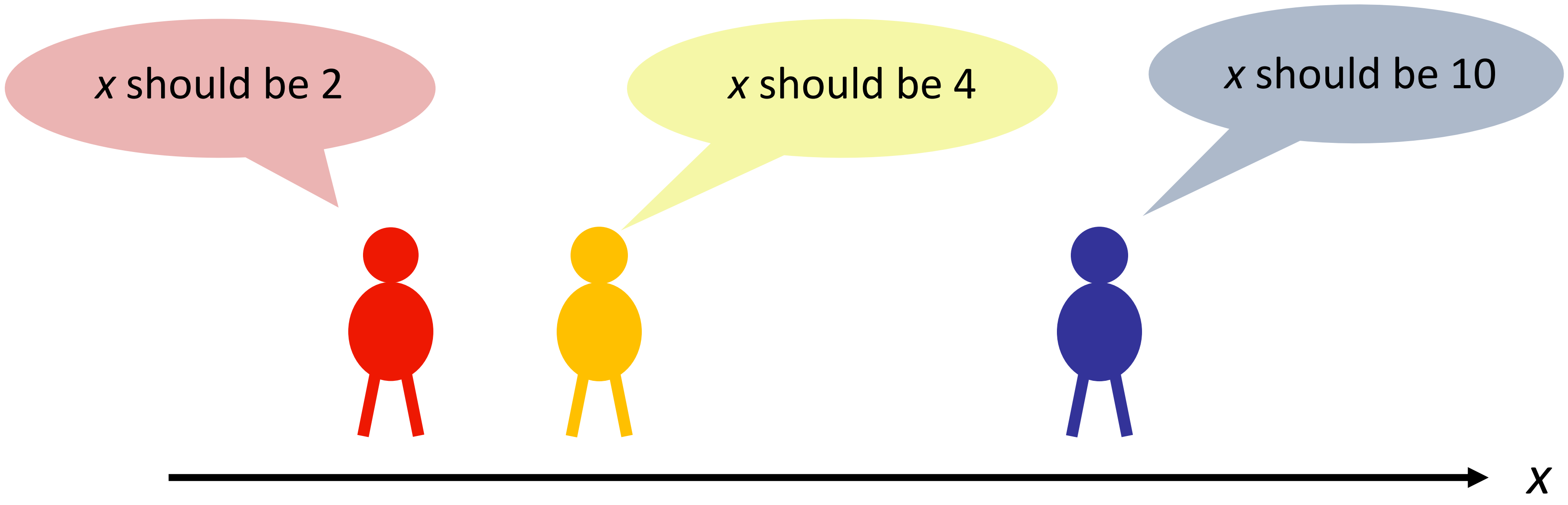
is as bad as



using ***x*** gallons
of gasoline

*How to determine **x**?*

One Approach: Let’s Vote!

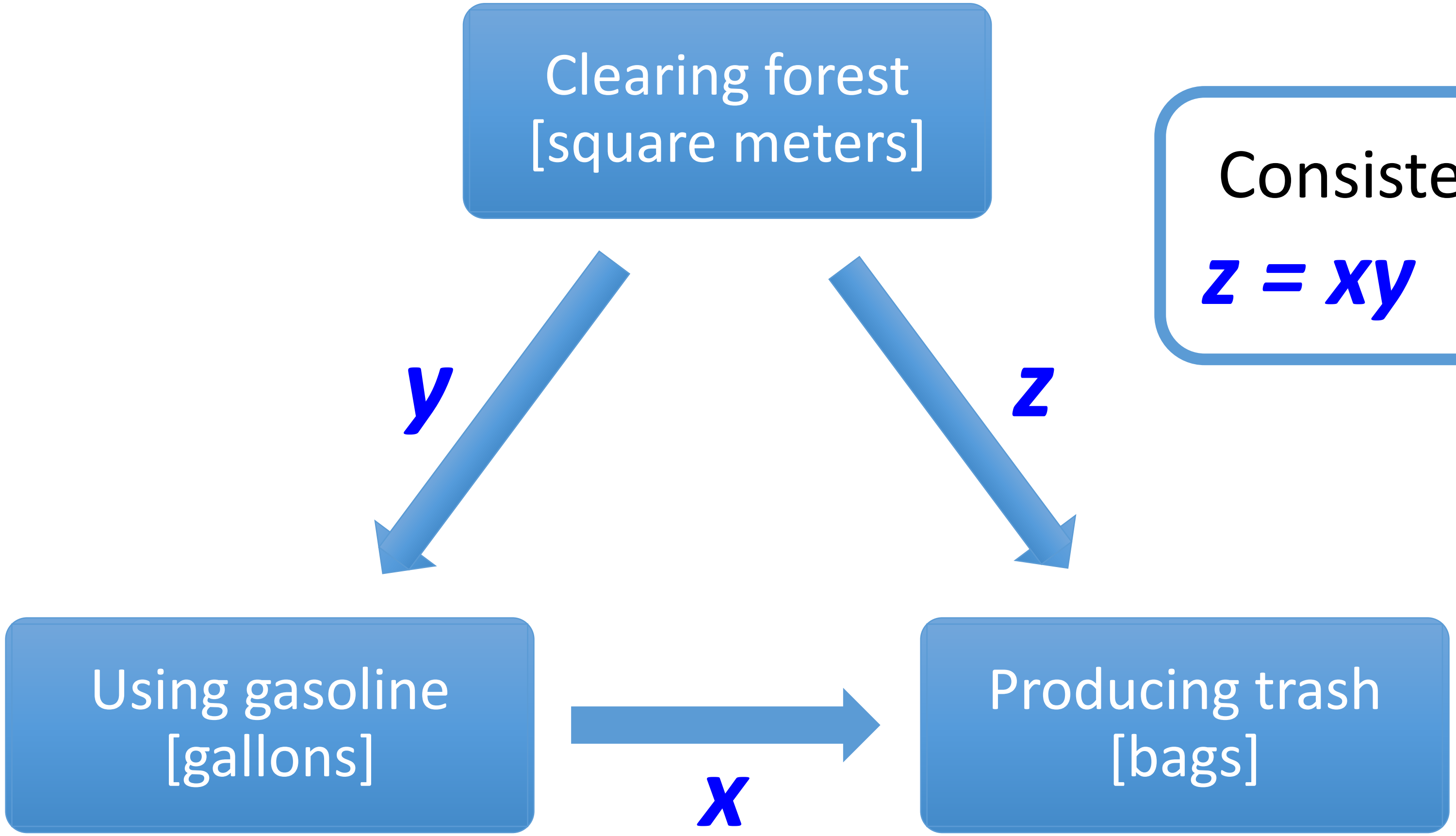


- What should the outcome be…?
  - Average? Median?

1  = x 

- Assuming that preferences are single-peaked, selecting the median is strategy-proof and has other desirable social choice-theoretic properties

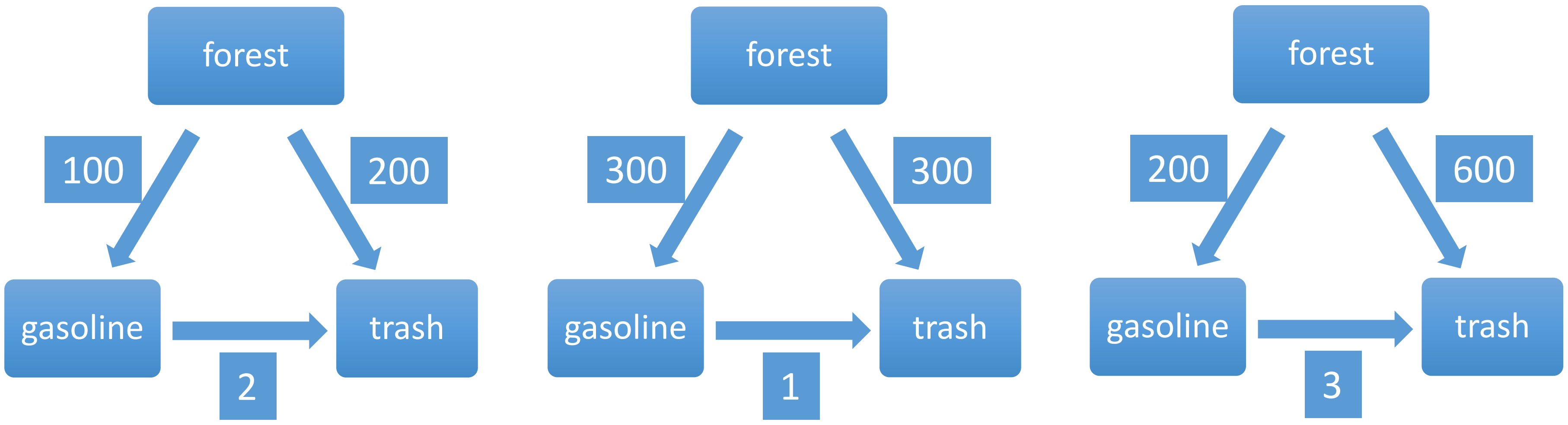
Consistency of tradeoffs



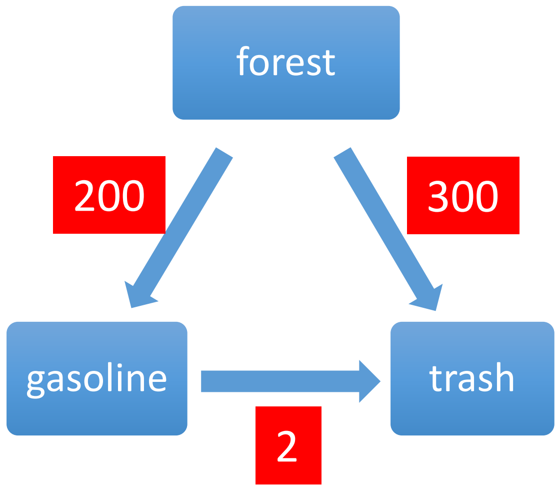
Consistency:

$z = xy$

A paradox

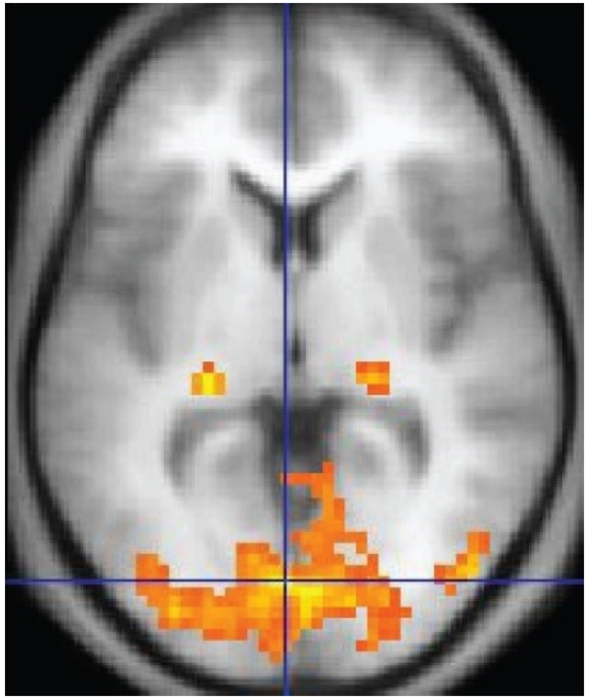


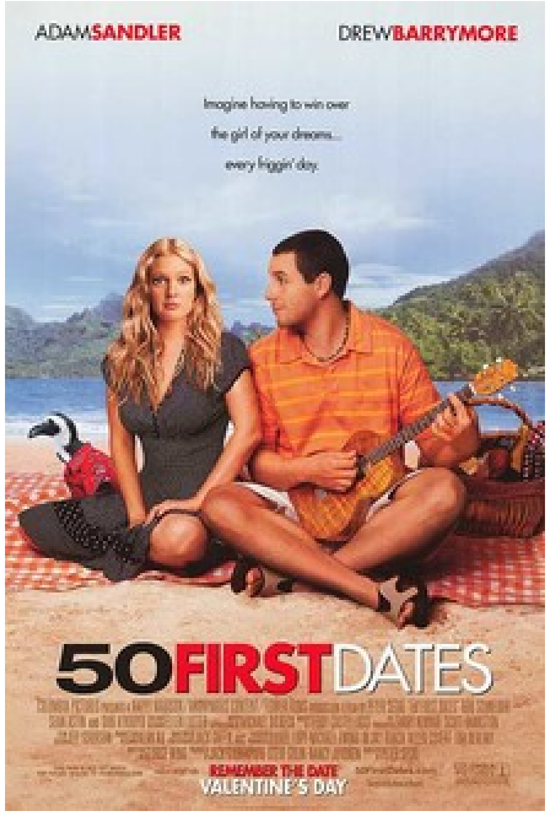
Just taking medians pairwise results in inconsistency



PART II. What should you do if…

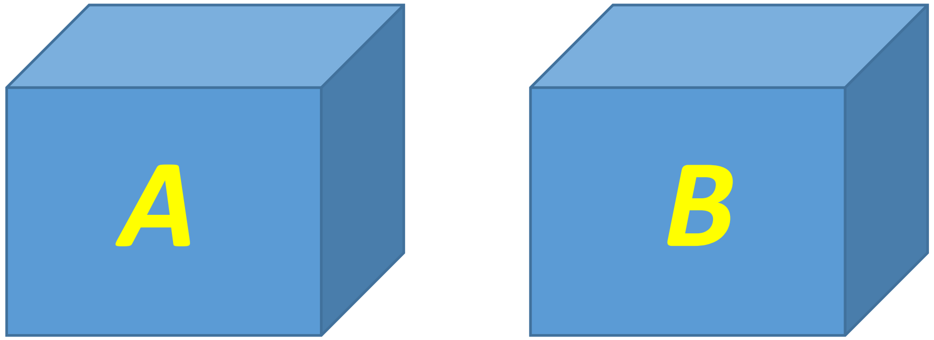
- … you knew *others could read your code*?
- … you knew *you were facing someone running the same code*?
- … you knew *you had been in the same situation before but can't possibly remember what you did*?



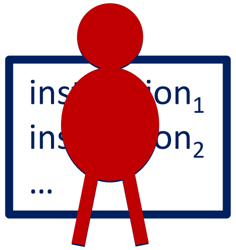


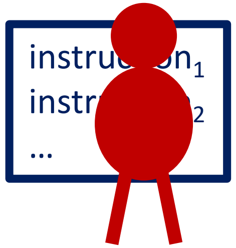
Newcomb's Demon

- 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





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

- 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?

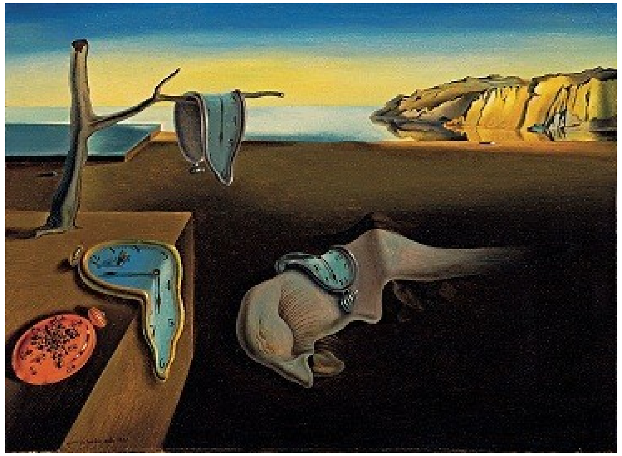
related to working paper
[Oesterheld, Demski, C.]

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	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

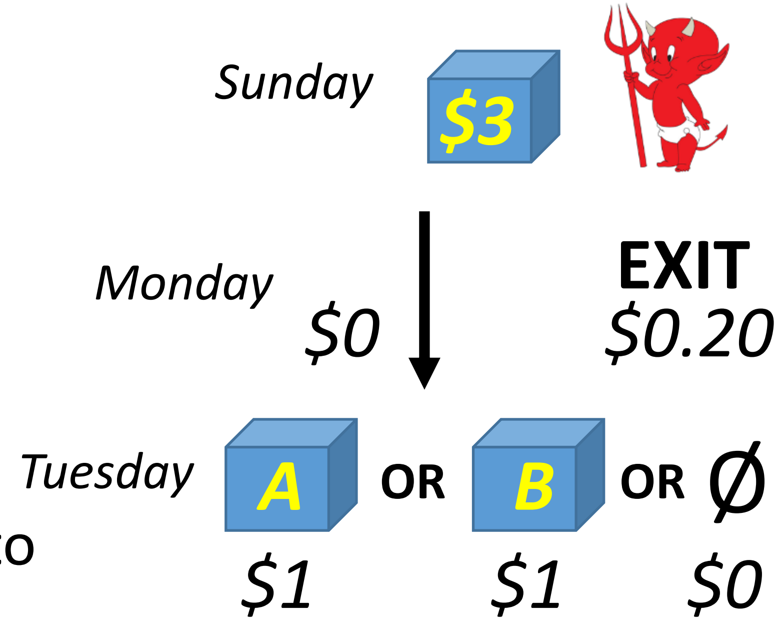
- *Evidential Decision Theory (EDT)*: When considering how to make a decision, consider how happy you expect to be conditional on taking each option and choose an option that maximizes that

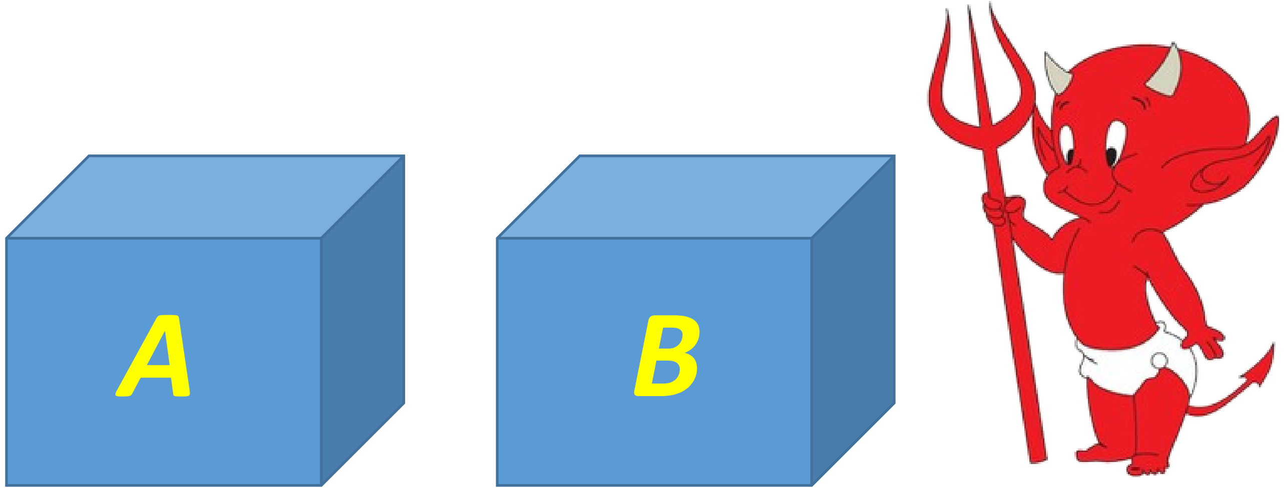
- *Causal Decision Theory (CDT)*: Your decision should focus on what you causally affect

Turning causal decision theorists into money pumps

[Oesterheld and C., Phil. Quarterly]

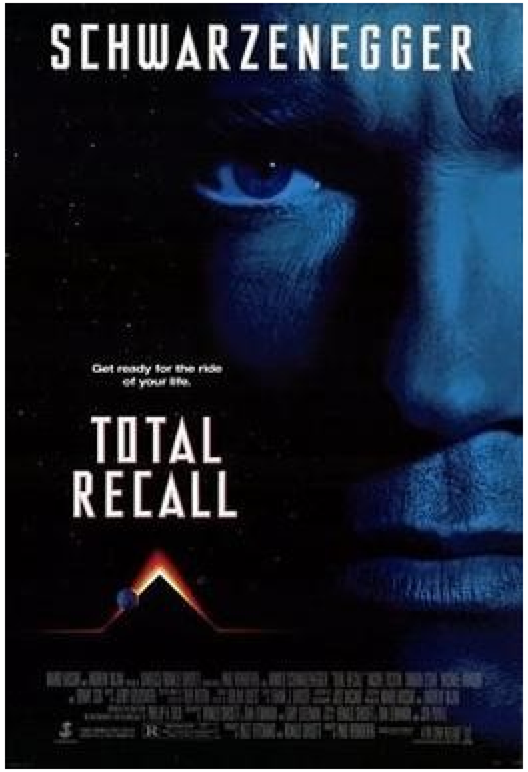
- **Adversarial Offer:**
- Demon (really, any good predictor) put $3 into each box it predicted you would not choose
- Each box costs $1 to open; can open at most one
- Demon 75% accurate (you have no access to randomization)
- CDT will choose one box, *knowing that it will regret doing so*
- Can add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





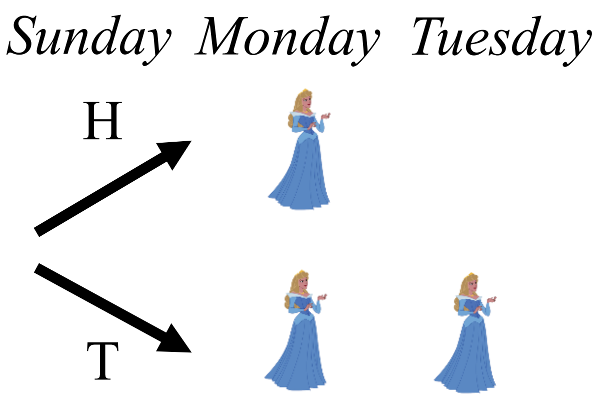
Imperfect recall

- 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'00]

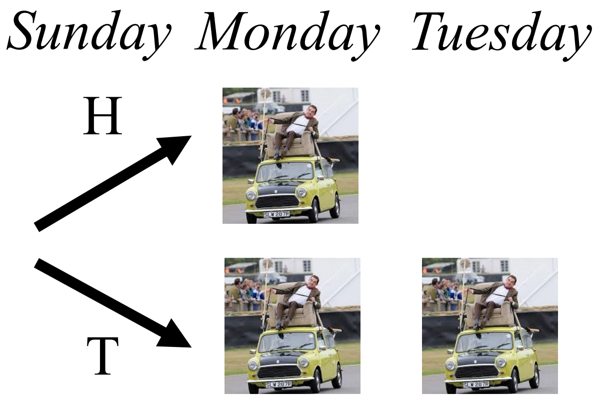
- There is a participant in a study (call her Sleeping Beauty)
- On Sunday, she is given drugs to fall asleep
- A coin is tossed (H or T)
- If H, she is awoken on Monday, then made to sleep again
- If T, she is awoken Monday, made to sleep again, then **again** awoken on Tuesday
- Due to drugs she **cannot 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?



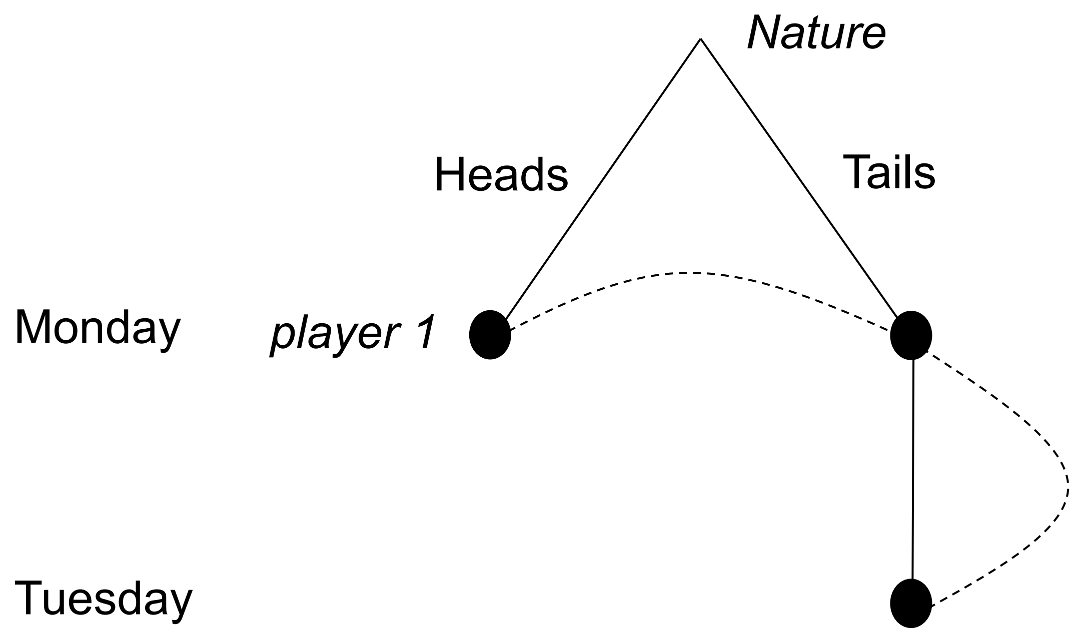
don't do this at home / without IRB approval…

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?

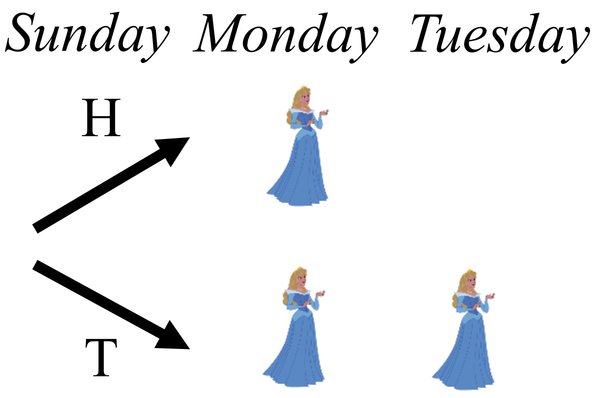


Information structure

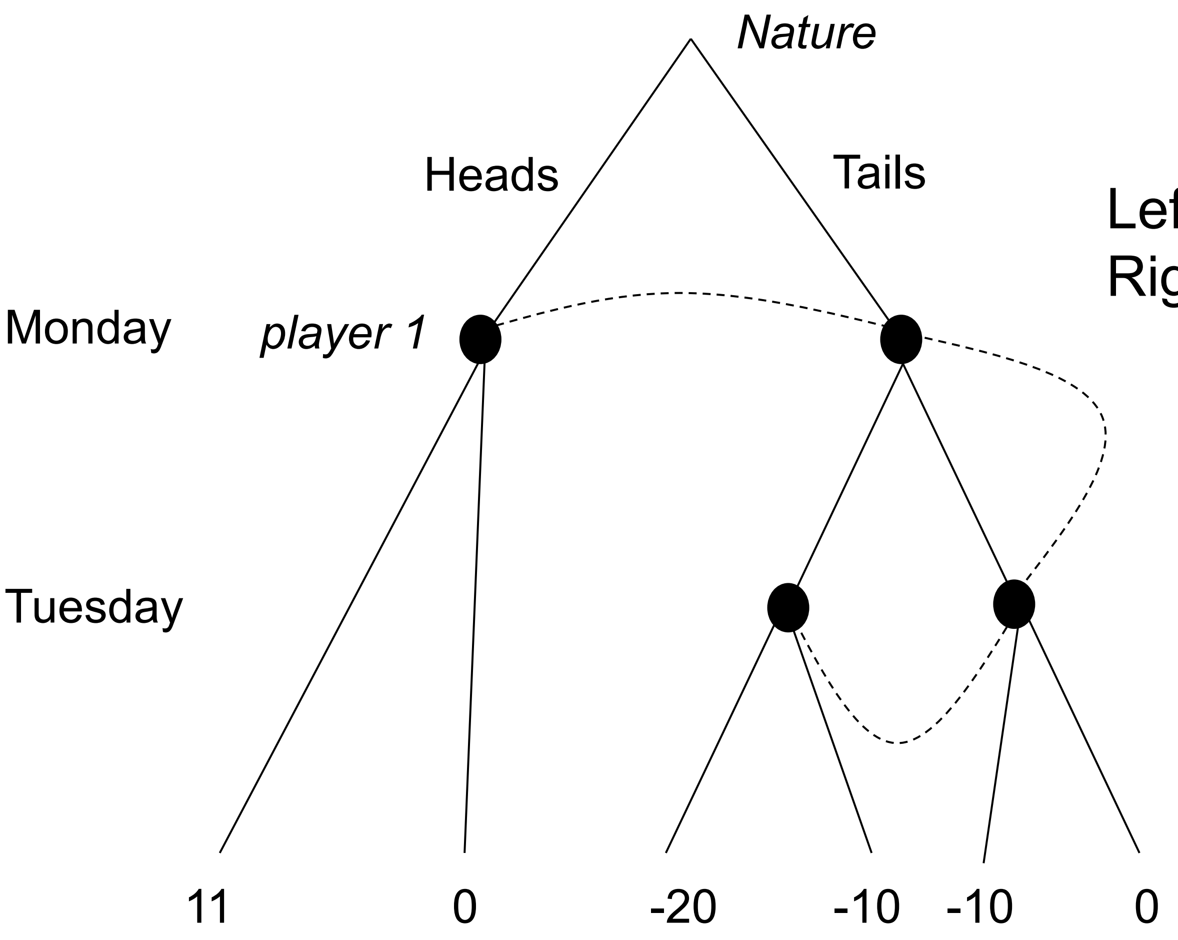


Taking advantage of a Halfer [Hitchcock'04]

- 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)



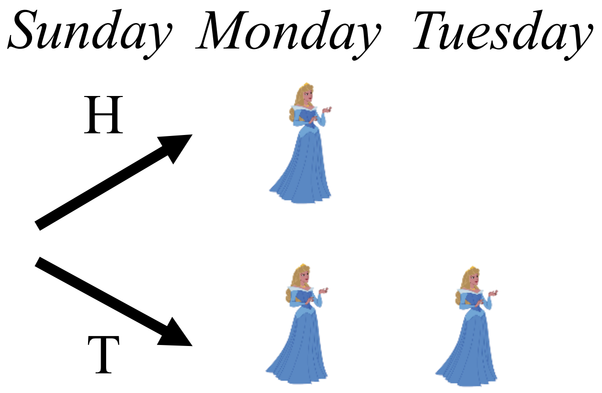
The betting game



Left=accept,
Right= decline

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. ½, it's Heads; if I accept, I will end up with 11. With prob. ½, 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. ½, it's Heads; if I accept, it will pay off 11. With prob. ½, 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
  - [C., Synthese'15]
  - … though we've argued against CDT in other work [Oesterheld & C, Phil. Quarterly'21]



Dutch book against EDT [C. 2015]

- Modified version of Sleeping Beauty where she wakes up in rooms of various colors

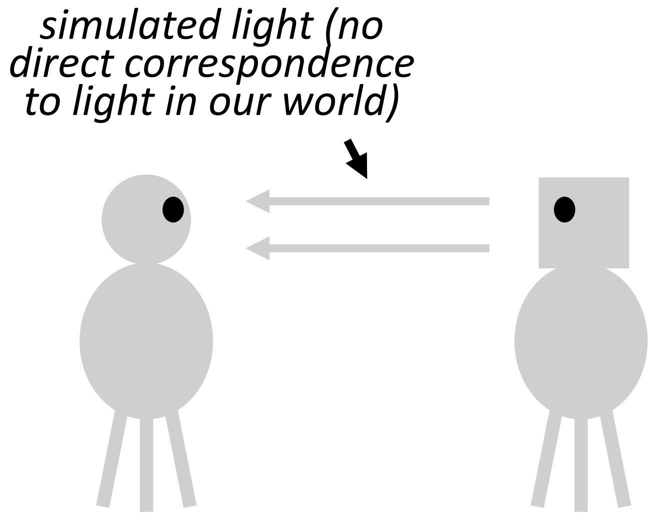
	WG (1/4)	WO (1/4)	BO (1/4)	BG (1/4)
Monday	white	white	black	black
Tuesday	grey	black	white	grey

Fig. 3 Sequences of coin tosses and corresponding room colors, as well as their probabilities, in the WBG Sleeping Beauty variant.

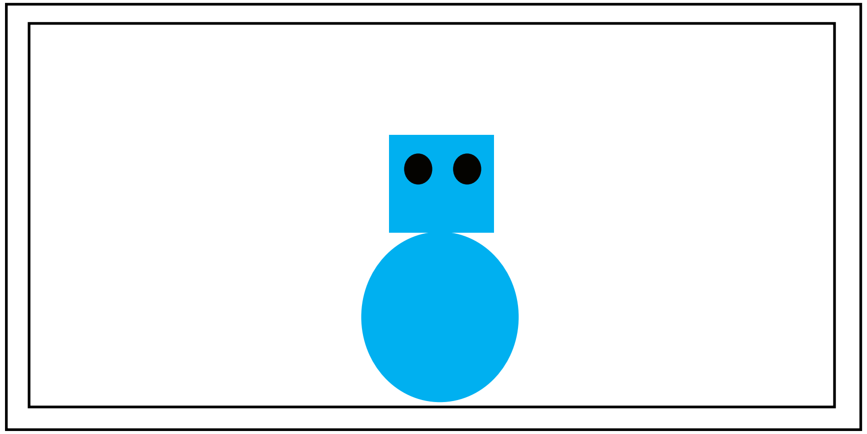
	WG (1/4)	WO (1/4)	BO (1/4)	BG (1/4)
Sunday	bet 1: 22	bet 1: -20	bet 1: -20	bet 1: 22
Monday	bet 2: -24	bet 2: 9	bet 2: 9	bet 2: -24
Tuesday	no bet	bet 2: 9	bet 2: 9	no bet
total gain from accepting all bets	-2	-2	-2	-2

Fig. 4 The table shows which bet is offered when, as well as the net gain from accepting the bet in the corresponding possible world, for the Dutch book presented in this paper.

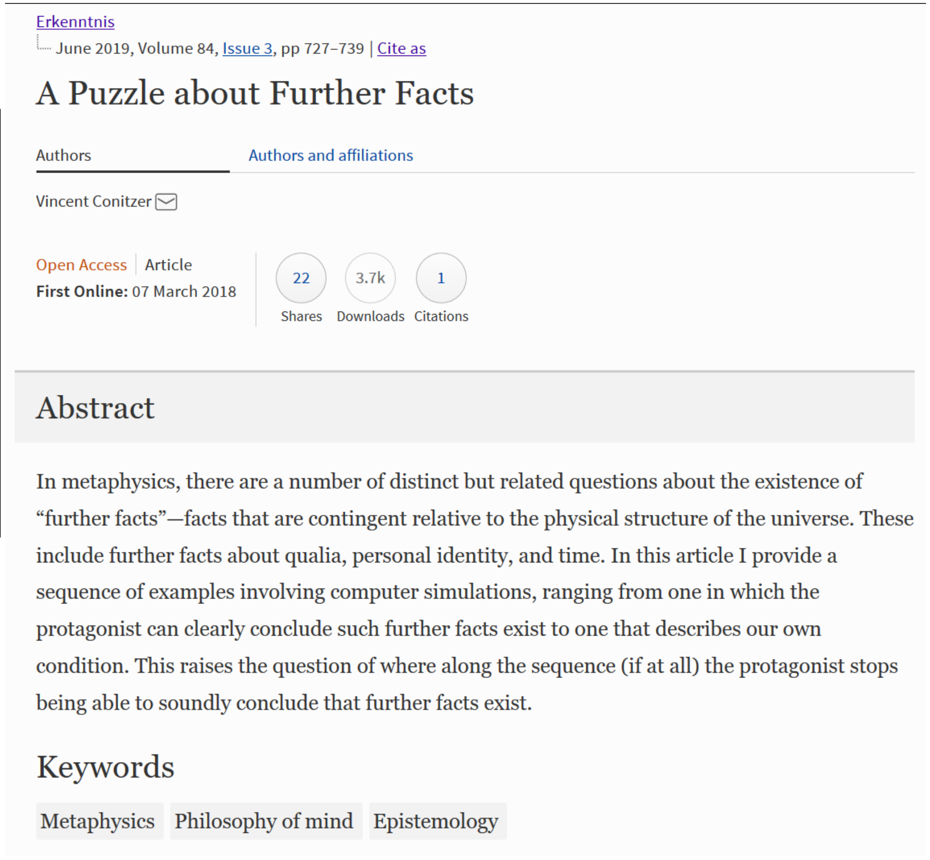
Philosophy of “being present” somewhere, sometime



1: world with creatures simulated on a computer



2: displayed perspective of one of the creatures


Erkenntnis
June 2019, Volume 84, Issue 3, pp 727–739 | Cite as

A Puzzle about Further Facts

Authors | Authors and affiliations

Vincent Conitzer

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

- 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

See also: [Hare 2007-2010, Valberg 2007, Hellie 2013, Merlo 2016, …]

- 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?

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 halfing… Shouldn't b depend on p...

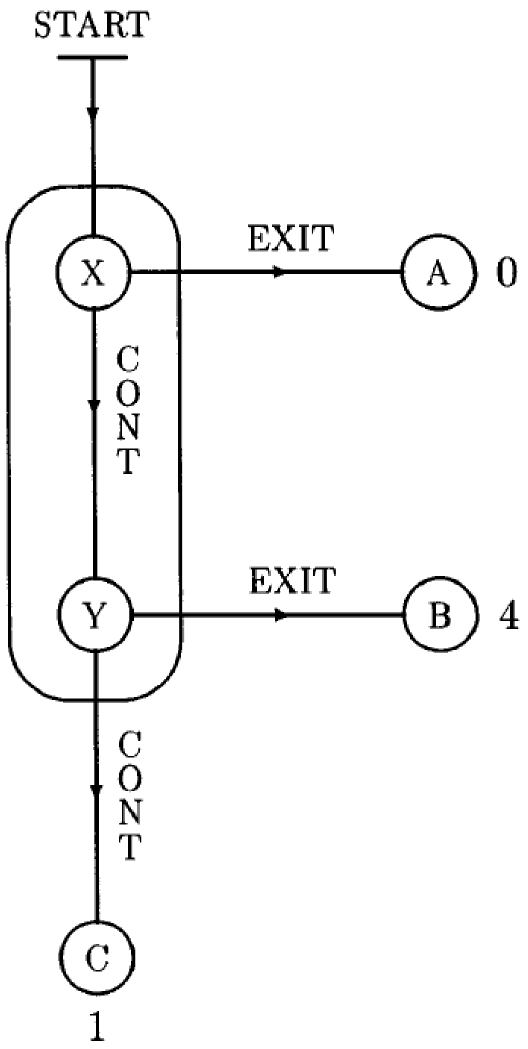


FIG. 1. The absent-minded driver problem.

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

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

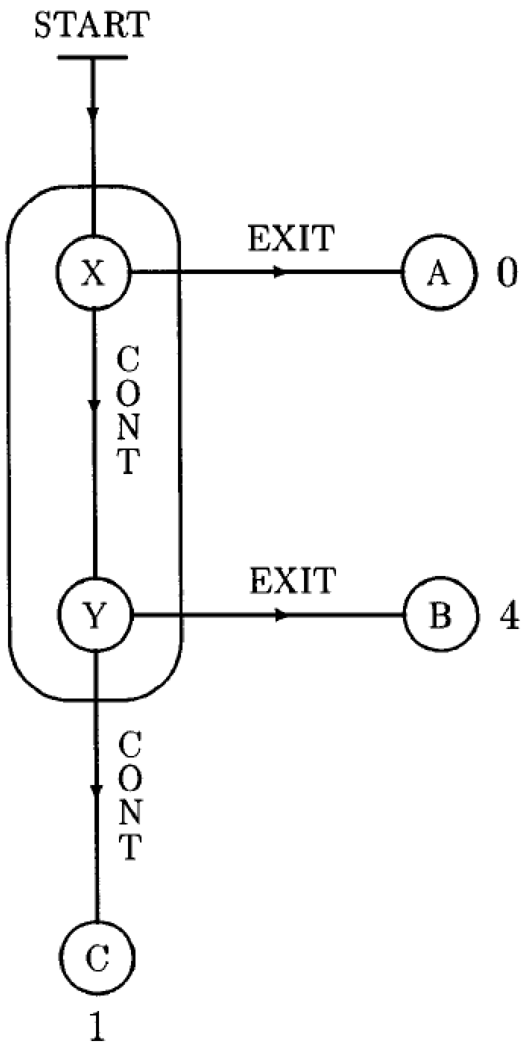
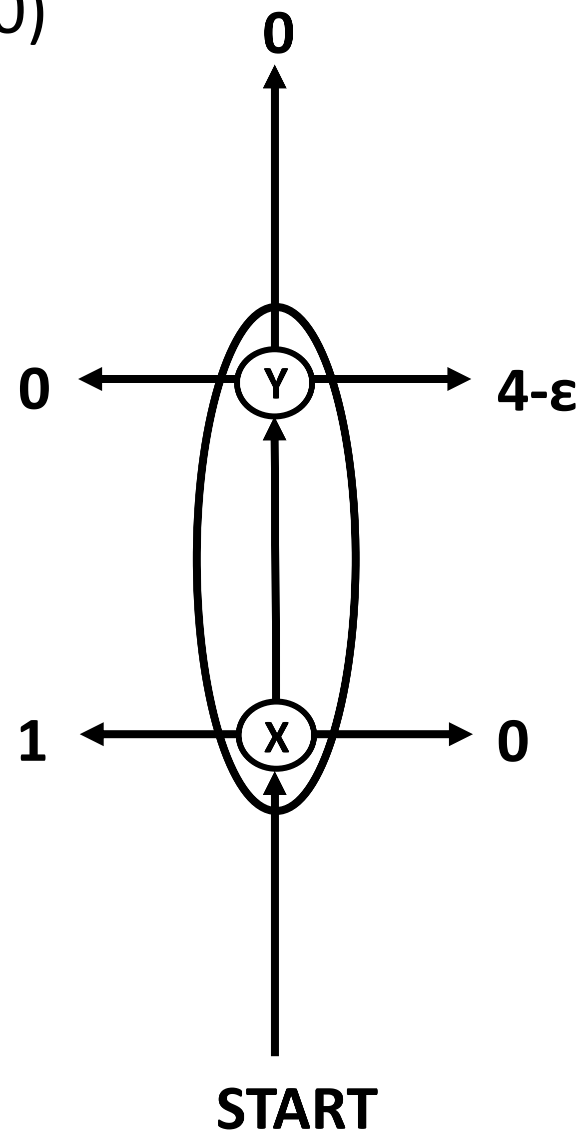


FIG. 1. The absent-minded driver problem.

Image from Aumann, Hart, Perry 1997

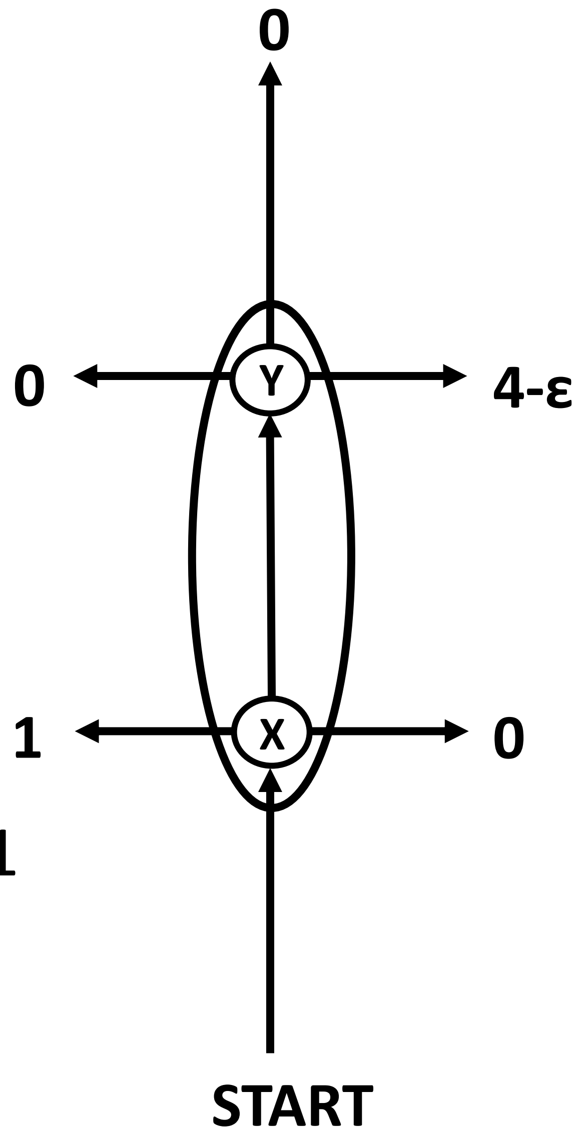
A challenging example for the evidential decision theorist

- Optimal strategy to commit to is to just go left: $(p_l, p_s, p_r) = (1, 0, 0)$
- If you're at an intersection, what does EDT say you should do?
- When considering $(p_l, p_s, p_r) = (1, 0, 0)$, you presumably expect to be at X and get 1 (really just need: no more than 1)
- When considering $(p_l, p_s, p_r) = (0, ½, ½)$, then say b is your subjective probability of being at Y
  - **Assume:** $b > 0$
  - **Assume:** b is not a function of ε
- So, expected utility: b*½*(4-ε) + (1-b)*¼*(4-ε) = 1+b-¼ε-¼bε
- For sufficiently small ε this is greater than 1
- Hence EDT suggests (0, ½, ½) over (1, 0, 0)!
- ... right? ... right?



A way for EDT to get the right answer (+SSA)

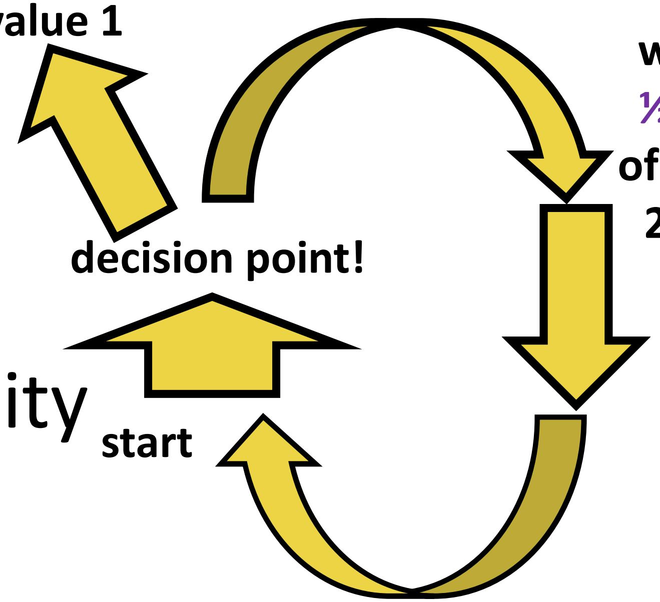
- Consider probabilities of **whole trajectories, plus where you are**, under strategy (0, ½, ½), **in a halfing sort of way**
- P(XY(4-ε), @X) = P(XY(4-ε)) * P(@X|XY(4-ε)) = ¼ * ½
- P(XY(4-ε), @Y) = P(XY(4-ε)) * P(@Y|XY(4-ε)) = ¼ * ½
- Any other trajectory with positive probability gives payoff 0
- So expected utility is 2 * ¼ * ½ * (4-ε) = 1- ε/4, which is worse than 1, so EDT gets the right answer
- *What just happened?*
- Under this way of reasoning, if you tell me that I'm at X, it's **more likely** that I'm on trajectory X(0) than on one of the XY ones
- P(XY(4-ε), @X) = ¼ * ½ ; P(XY(0), @X) = ¼ * ½ ; P(X(0), @X) = ½ * 1
- So P(X(0) | @X) = ½ / (½ + ¼) = 2/3 (**not** 1/2)
- Previous slide had **hidden assumption**: *where I am carries no information about my **future** coin tosses*



Making decisions with imperfect recall

[cf. absentminded driver problem: PR97, AHP97]

- Optimal strategy without recall: go Right with probability 5/8. *(Outside view.)* Follow that.
- You arrive at decision point. What is the probability that you're there for the first time? *(Inside view.)*
- **Thirder:** in expectation 1 first awakening, and (1/2)(5/8)(16/25) / (1-(5/8)(16/25)) = 1/3 later awakenings, so probability of first time = 1/(4/3) = ¾
- Going Left gives 1 and going Right gives (1/2)(3/4)(2) + ((1/2)(3/4)+(1/4))(16/25)(3/8) / (1-(5/8)(16/25)) = 1
- **Theorem.** This is always true!
- … but can have other equilibria



discount factor (probability that game continues):
γ = 16/25 = .64

with:


Scott Emmons


Caspar Oesterheld


Andrew Critch


Stuart Russell

Fraction of time replicator dynamic finds **best** solution

A N	2	3	4	5
2	0.93	0.81	0.68	0.65
3	0.81	0.70	0.58	0.46
4	0.76	0.58	0.36	0.34
5	0.69	0.43	0.36	0.30

(a) RandomGame

A N	2	3	4	5
2	0.58	0.45	0.40	0.33
3	0.57	0.35	0.29	0.27
4	0.53	0.37	0.28	0.25
5	0.51	0.33	0.33	0.24

(b) CoordinationGame

N = #players (or #nodes)
A = #actions per player (or per node)

Functional Decision Theory

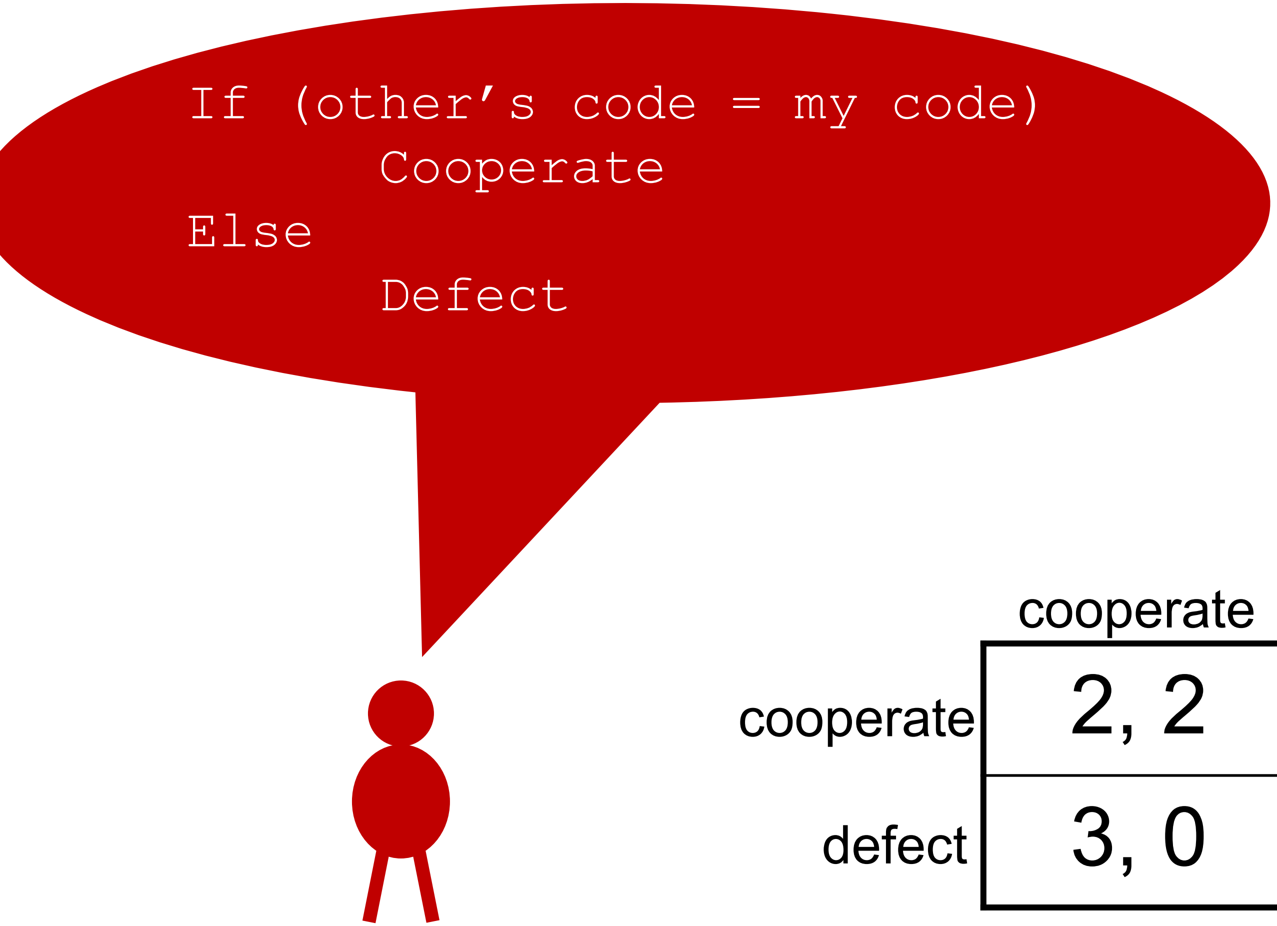
[Soares and Levinstein 2017; Yudkowsky and Soares 2017]

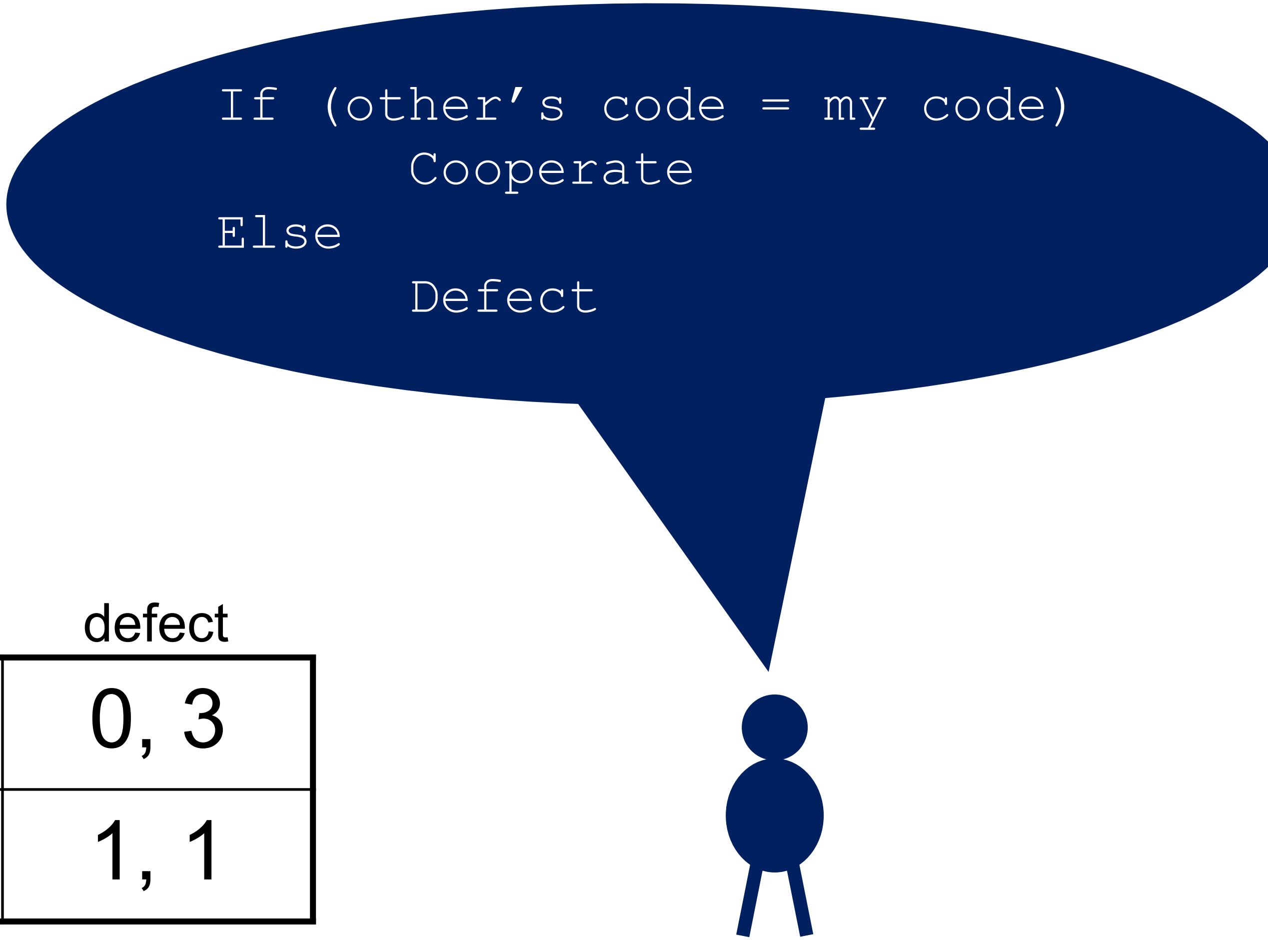
- One interpretation: *act as you would have precommitted to act*
- Avoids my EDT Dutch book (I think)
- … still one-boxes in Newcomb's problem
- … even one-boxes in Newcomb's problem **with transparent boxes**
- An odd example: Demon that will send you $1,000 if it believes you would otherwise destroy everything (worth -$1,000,000 to everyone)

- FDT says you should destroy everything, *even if you only find out that you are playing this game after the entity has already decided not to give you the money* (too-late extortion?)

Program equilibrium [Tennenholz 2004]

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





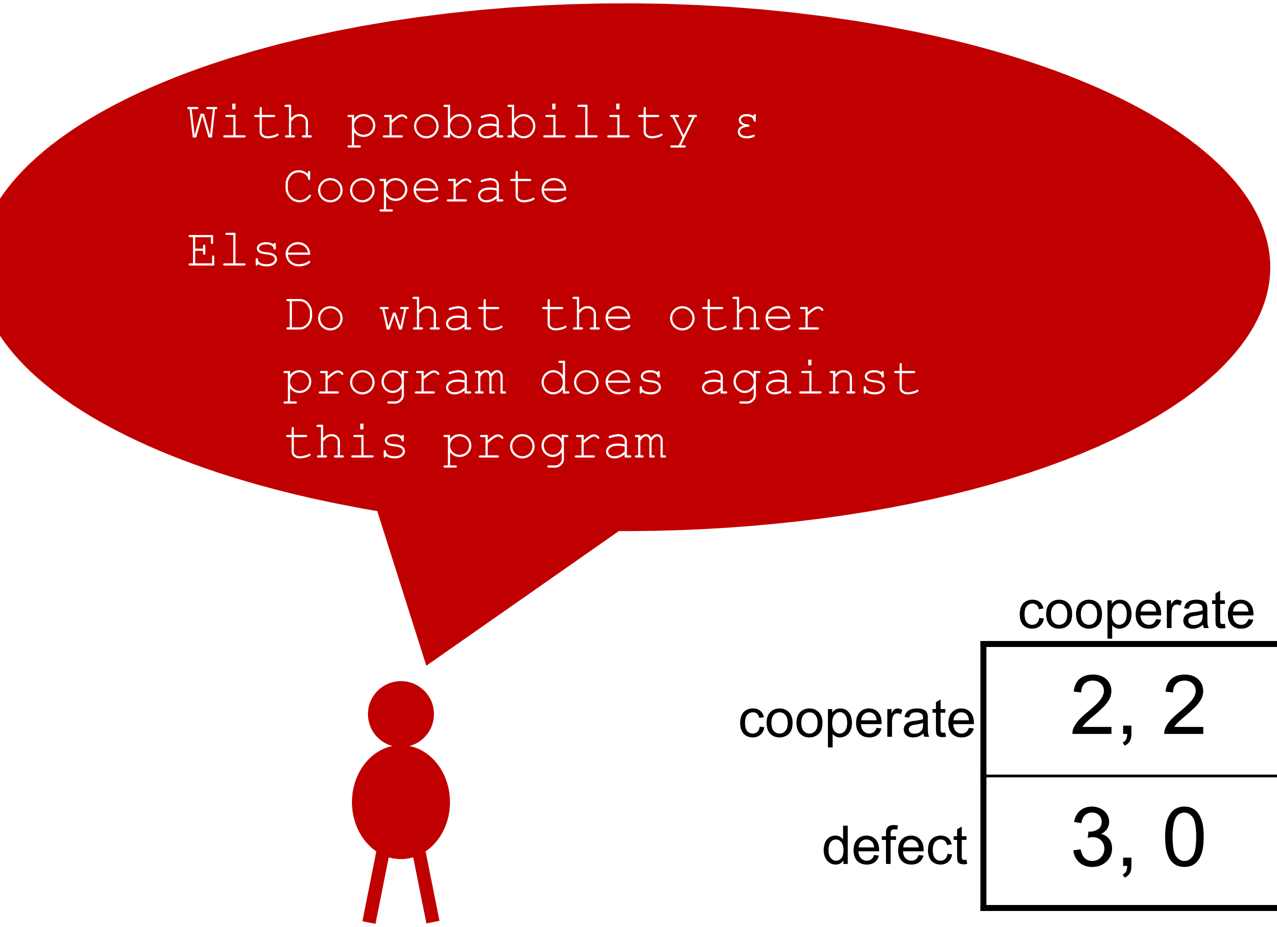
	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

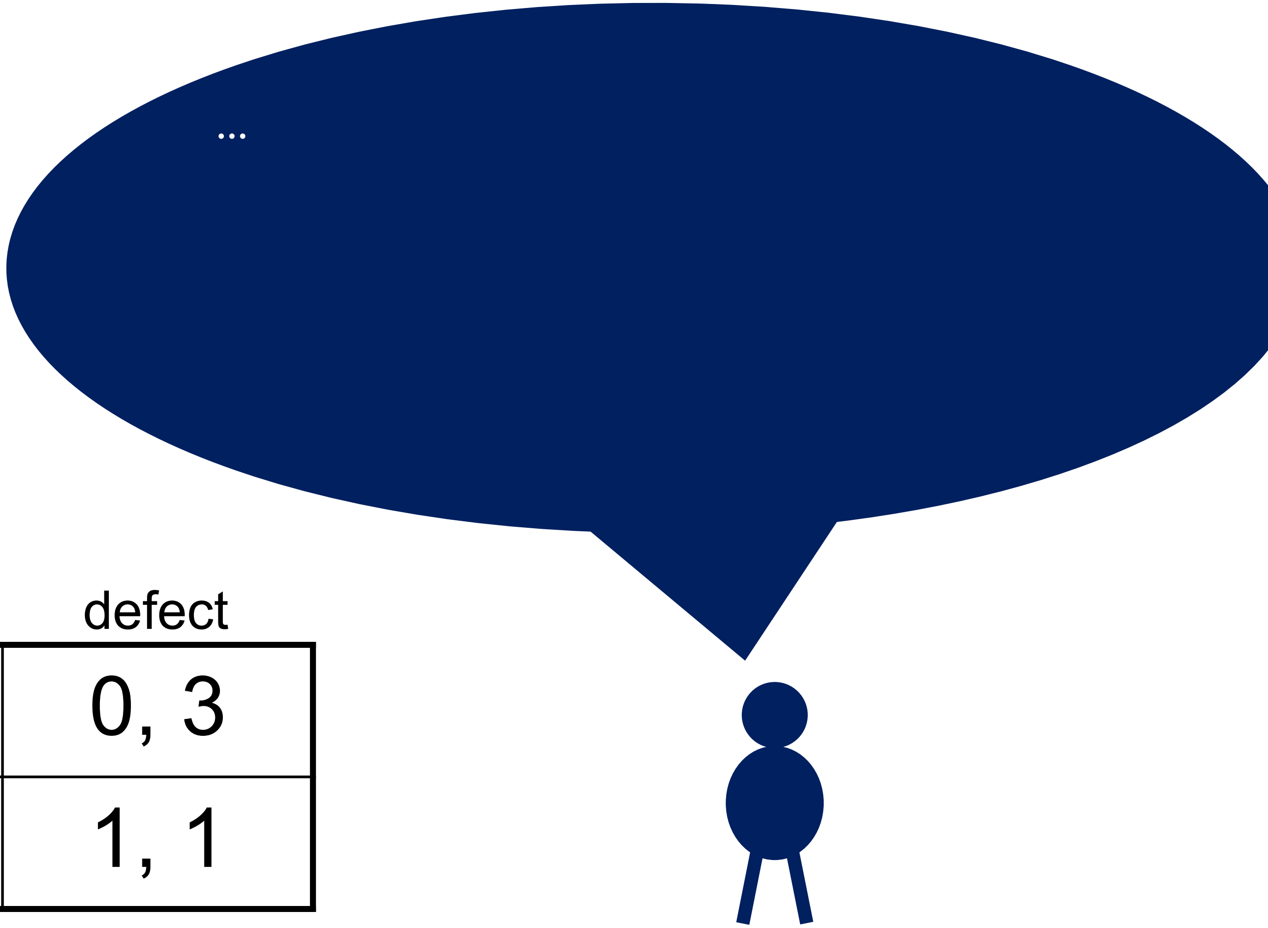
- See also: [Fortnow 2009, Kalai et al. 2010, Barasz et al. 2014, Critch 2016, Oesterheld 2018, …]

Robust program equilibrium [Oesterheld 2018]

- Can we make the equilibrium less fragile?

Caspar Oesterheld



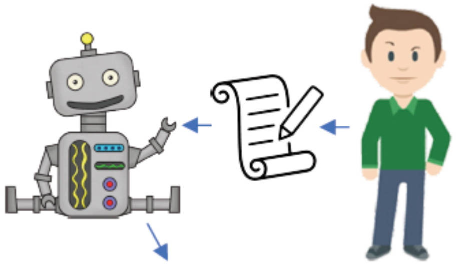


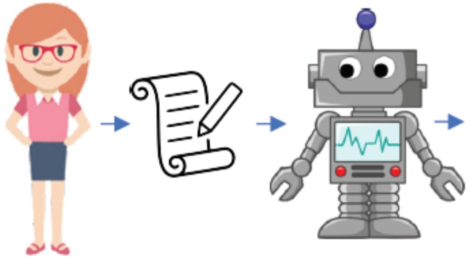
	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

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

Caspar Oesterheld

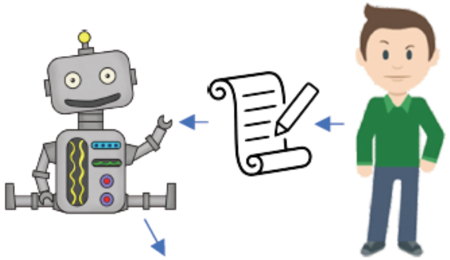
Delegated game playing

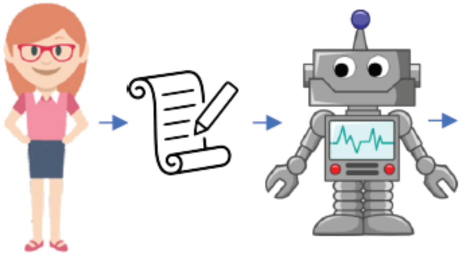




	DM	RM	DL	RL
DM	-5,-5	2,0	5,-5	5,-5
RM	0,2	1,1	5,-5	5,-5
DL	-5,5	-5,5	1,1	2,0
RL	-5,5	-5,5	0,2	1,1

- 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



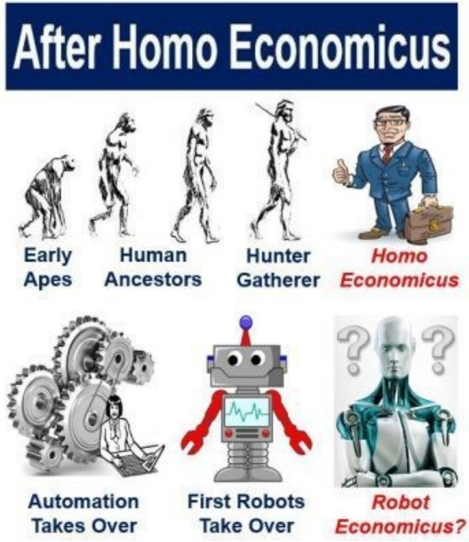


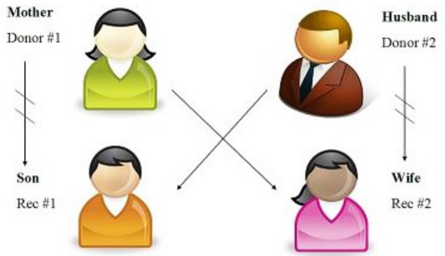
	DL	RL
DL	-5,-5 (1,1)	2,0 (2,0)
RL	0,2 (0,2)	1,1 (1,1)

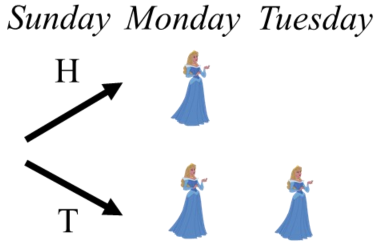
- Each player's contract says: Play this alternative game if the other player adopts an analogous contract.
- The games are essentially isomorphic.
  - DM ~ DL
  - RM ~ RL
- *Safe Pareto improvement* on the original game: outcome of new game is better for both players with certainty.

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
- 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







THANK YOU FOR YOUR ATTENTION!