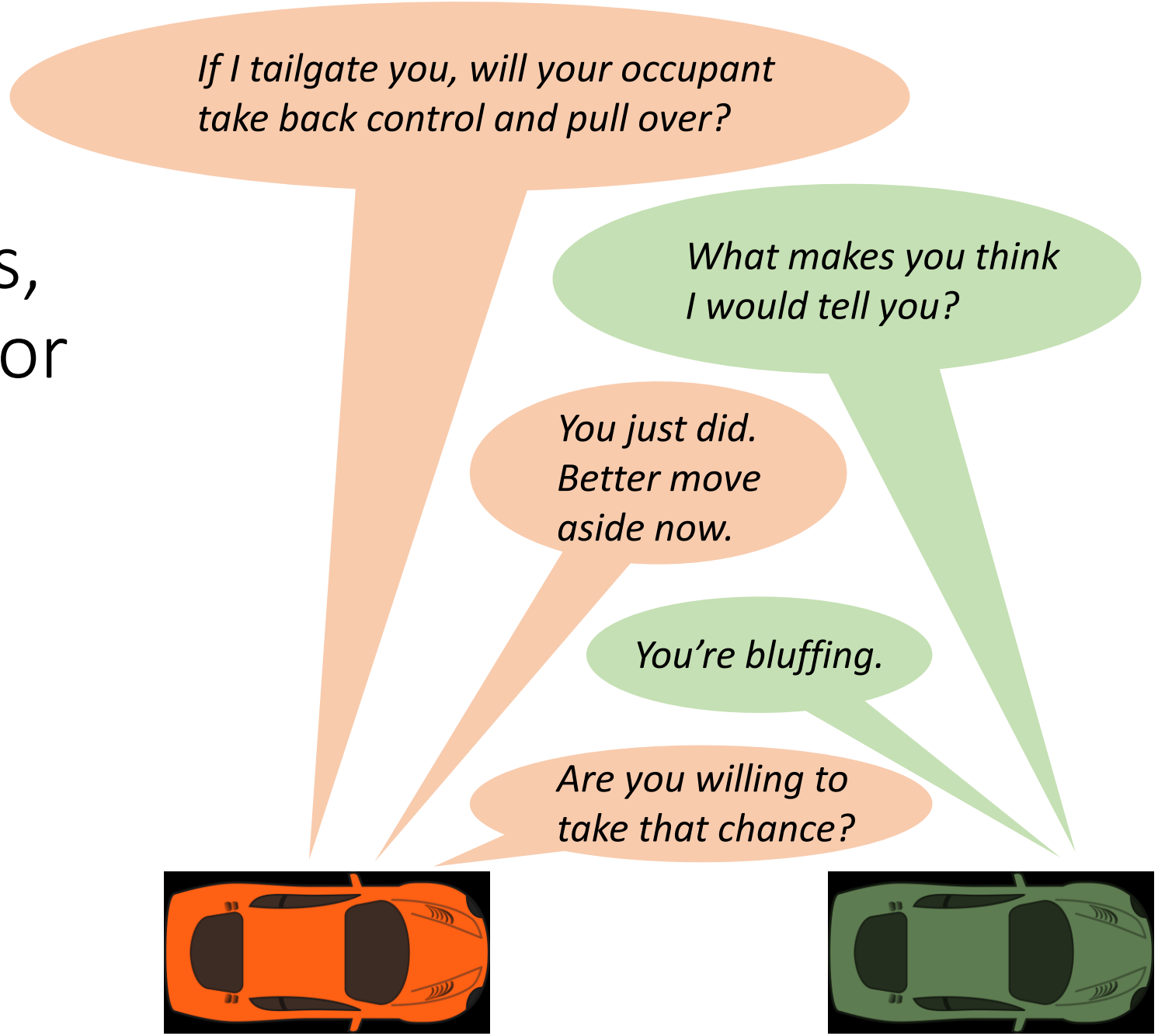
Designing Preferences, Beliefs, and Identities for Artificial Intelligence

Vincent Conitzer
(Duke University)



Russell and Norvig

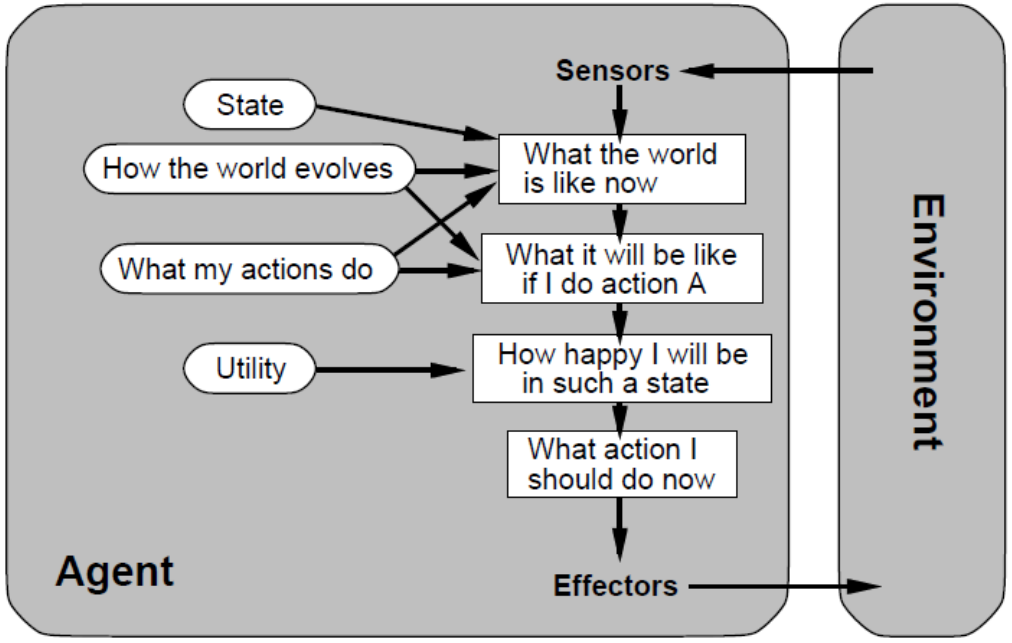
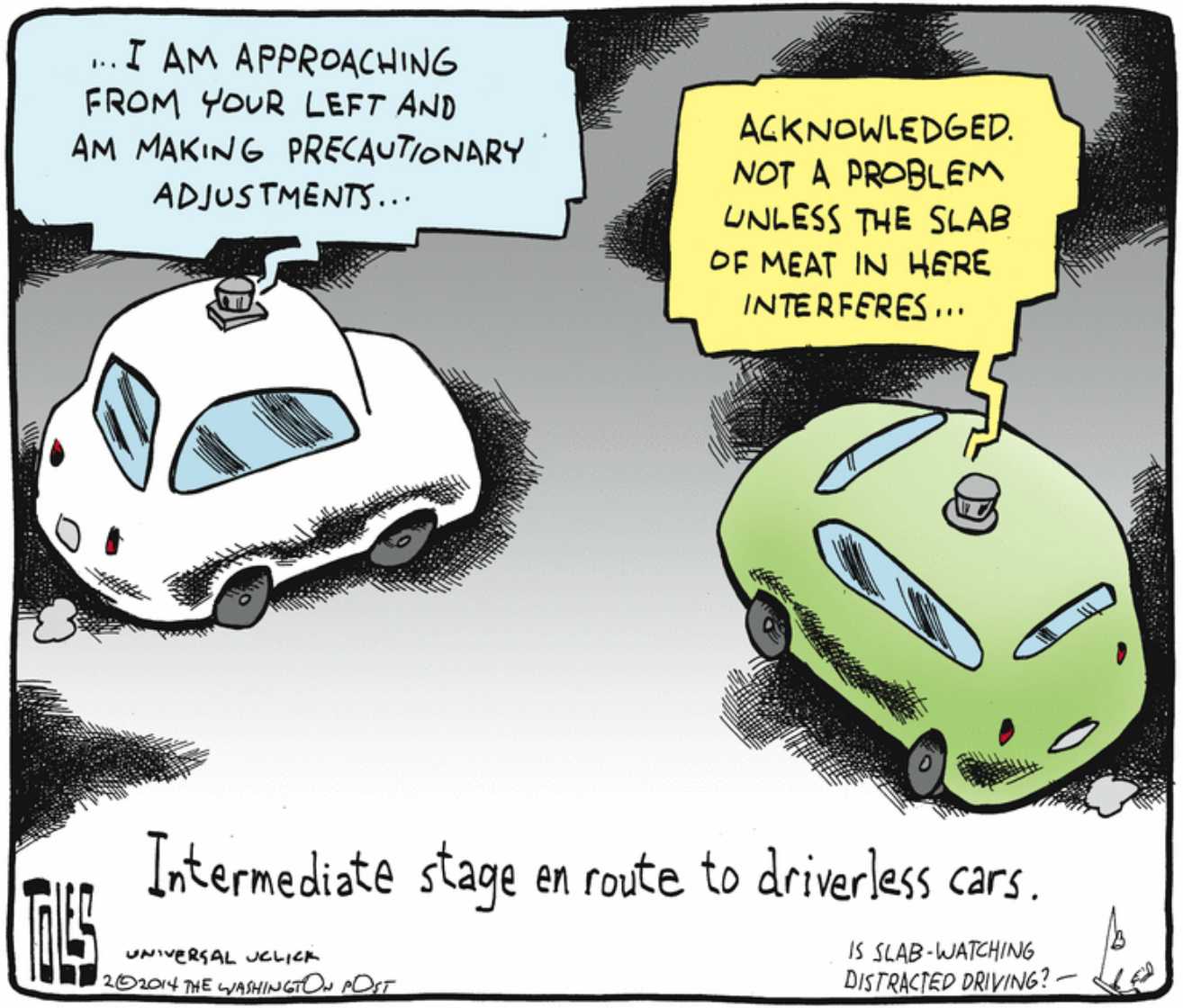


Figure 2.12 A complete utility-based agent.

"… 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."

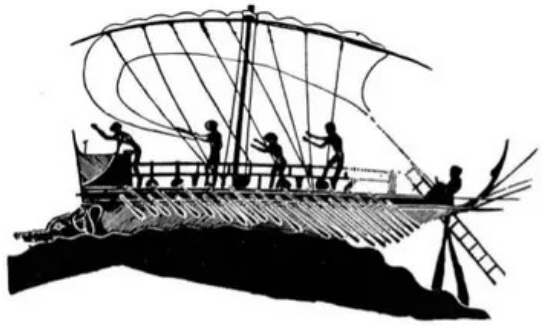
Example: network of self-driving cars



- 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; …)?

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

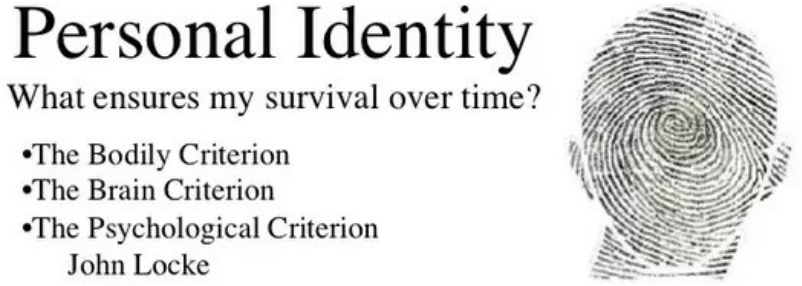
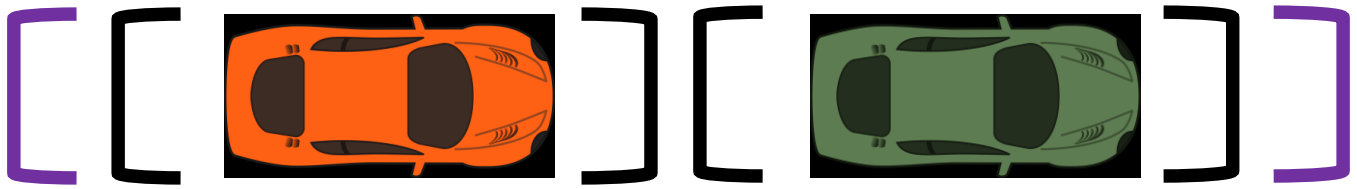


image from https://www.quora.com/What-solutions-are-there-for-the-Ship-of-Theseus-problem

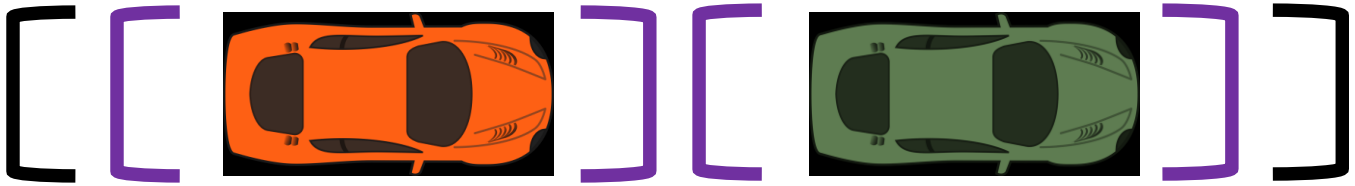
Splitting things up in different ways

beliefs

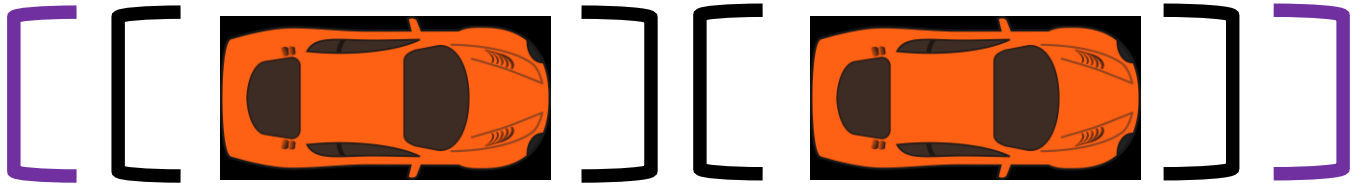
preferences



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)

$t = 1$ $t = 2$

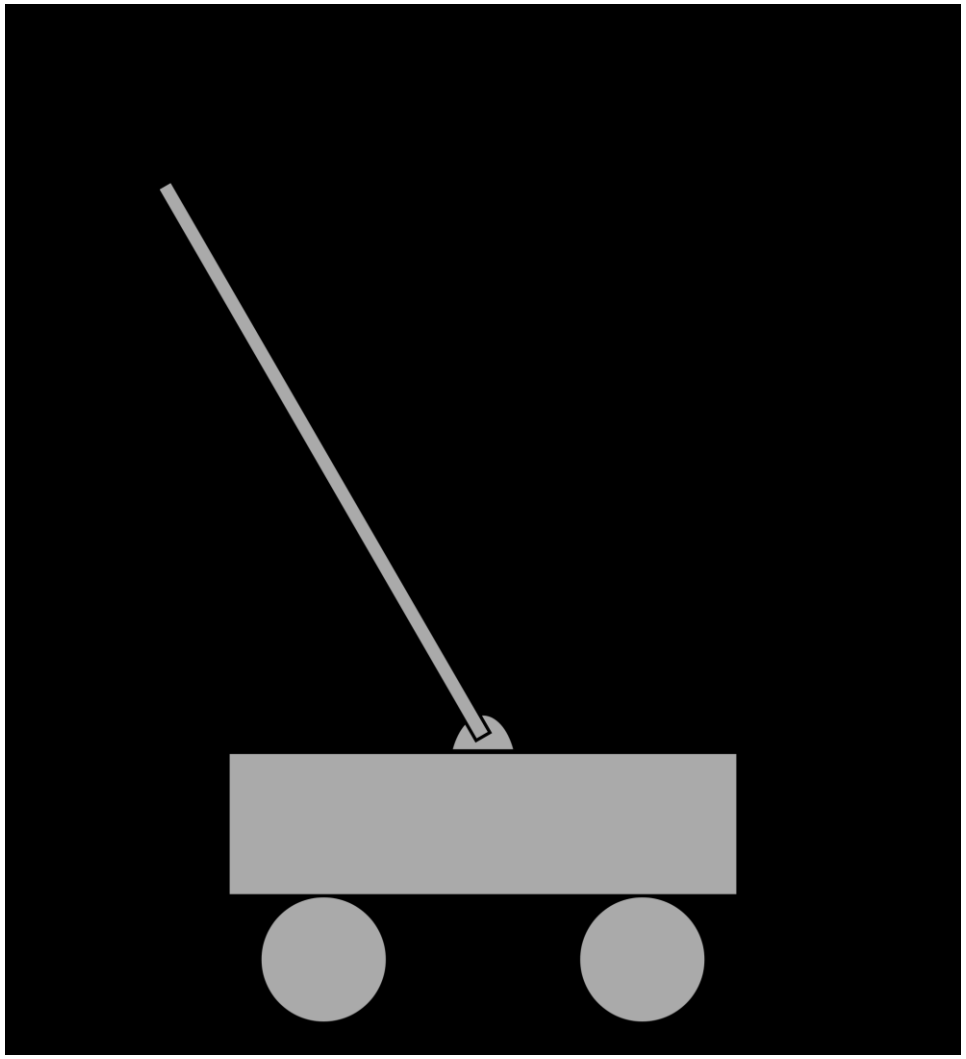
data is kept around but car acts on behalf of current owner

$t = 1$ $t = 2$

Outline

- Learning an objective from multiple people
  - Focus on moral reasoning
  - Use social choice theory
- Interlude: background on game theory (separate deck of slides)
- Decision and game-theoretic approaches to agent design
  - Improving equilibria by designing a few preferences
  - Imperfect recall and Sleeping Beauty
  - Causal and evidential decision theory (and others)
- 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.

…


AAAI /ACM Conference on
Artificial Intelligence, Ethics, and Society
Honolulu, Hawaii, USA
January 27-28, 2019
CALL FOR PAPERS

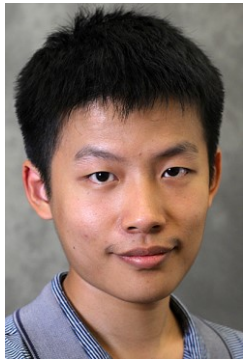
Moral Decision Making Frameworks for Artificial Intelligence

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

with:

Walter Sinnott-Armstrong

Jana Schaich Borg

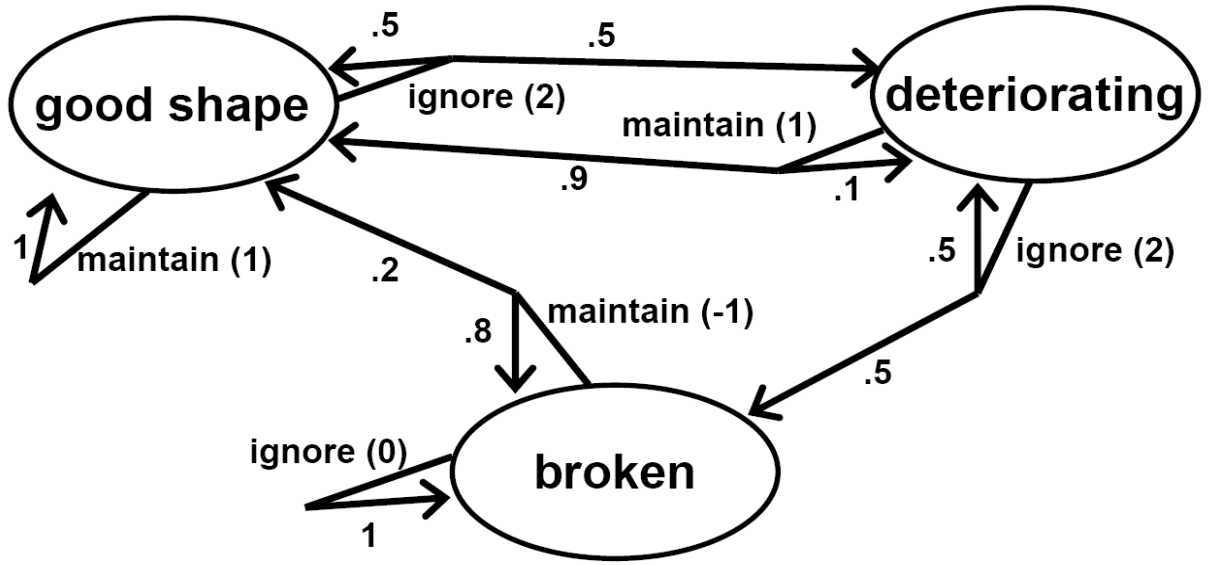


Yuan Deng

Max Kramer

The value of generally applicable frameworks for AI research

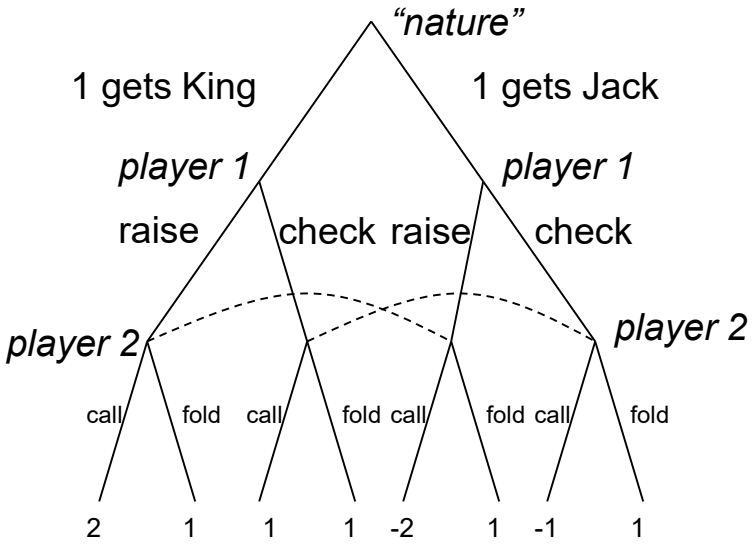
- Decision and game theory
- Example: Markov Decision Processes
- Can we have a **general** framework for moral reasoning?



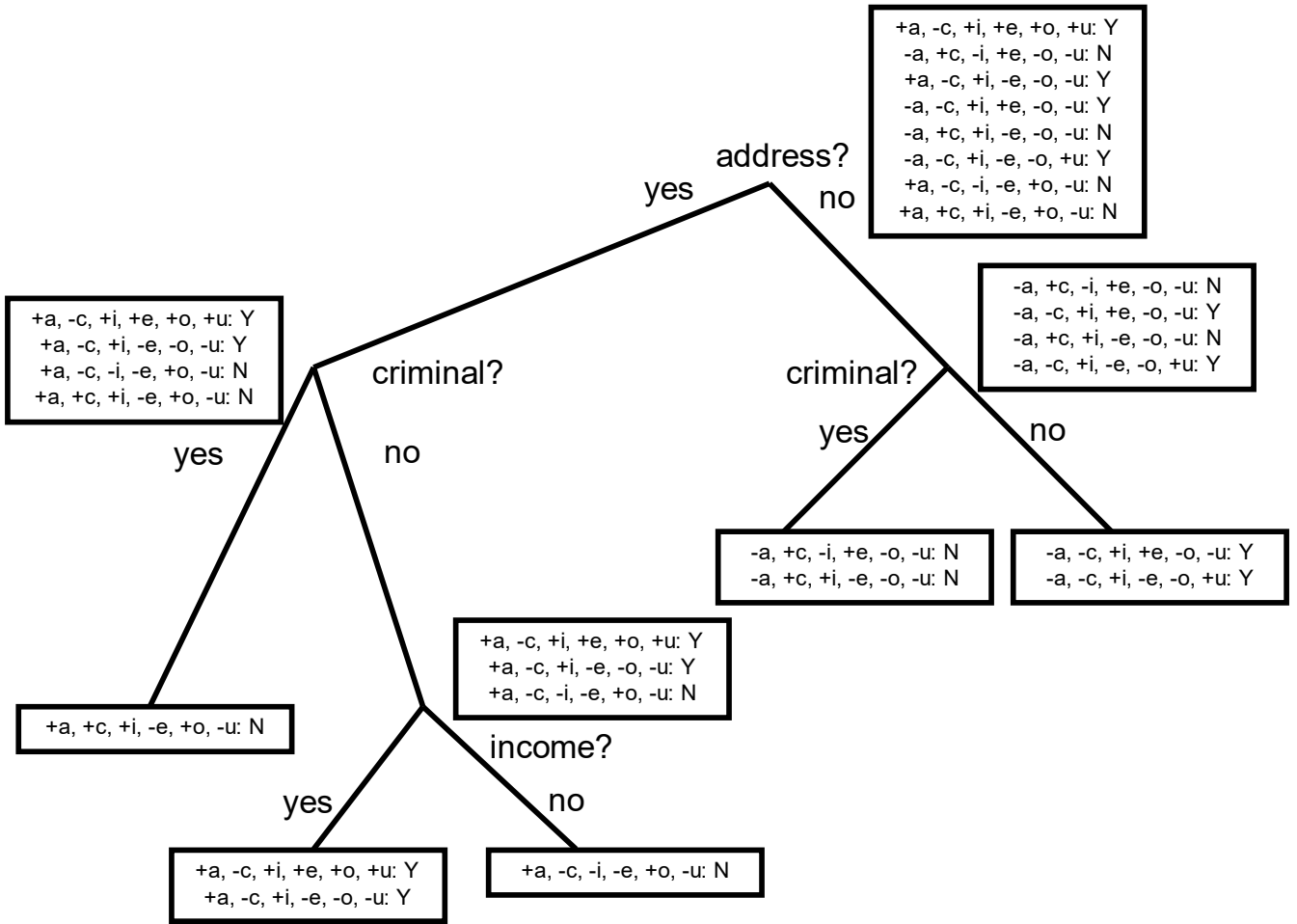
Two main approaches

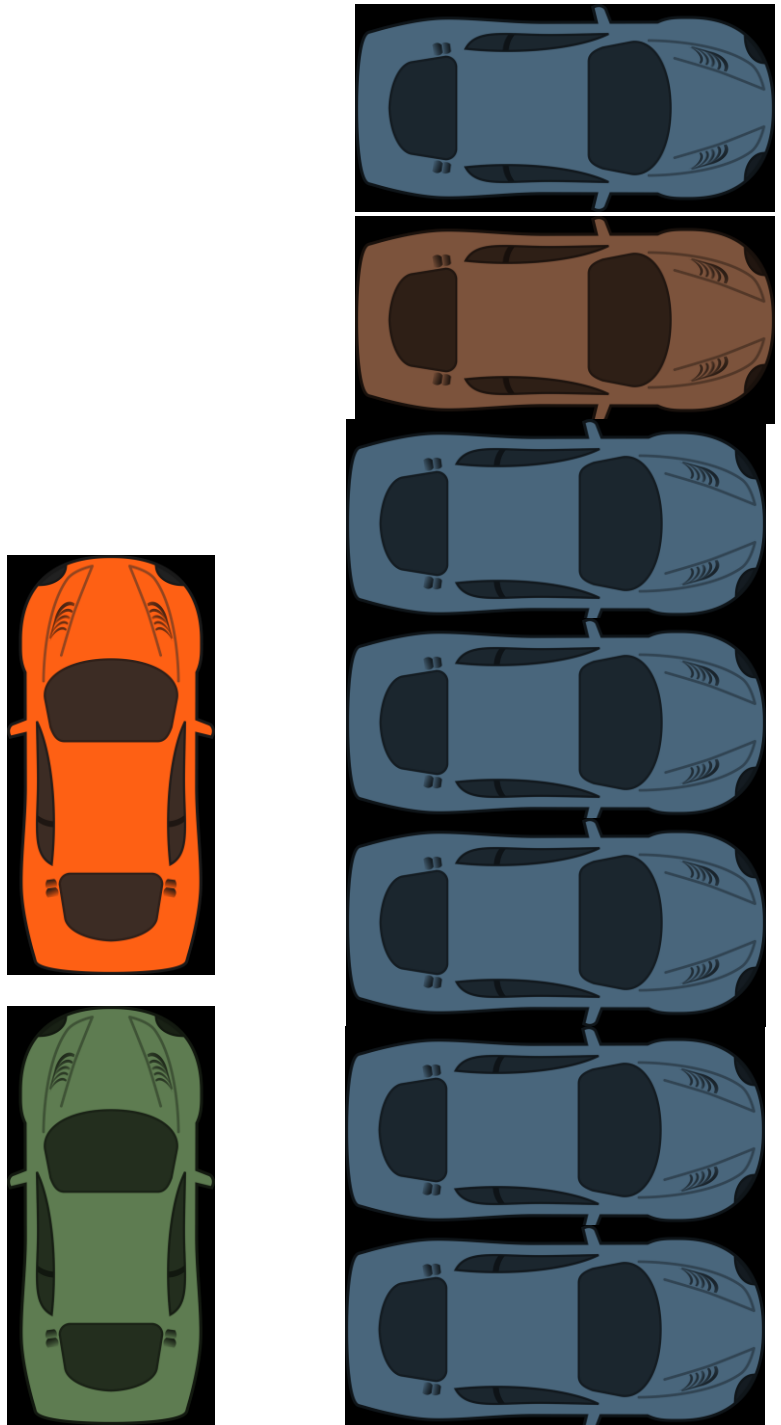
Cf. top-down vs. bottom-up distinction [Wallach and Allen 2008]

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



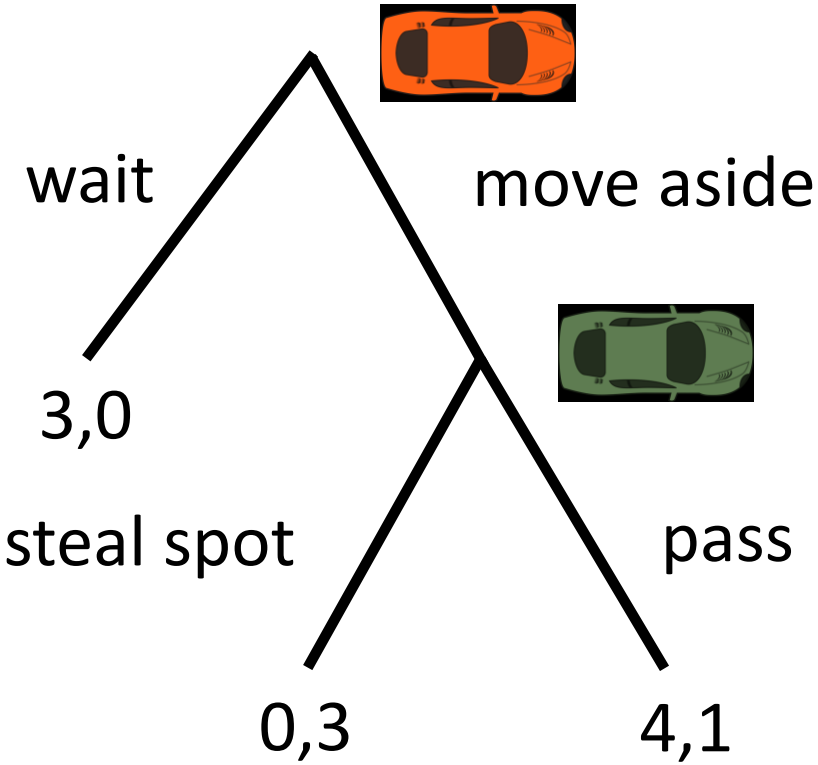
Generate data sets of human judgments, apply **machine learning**





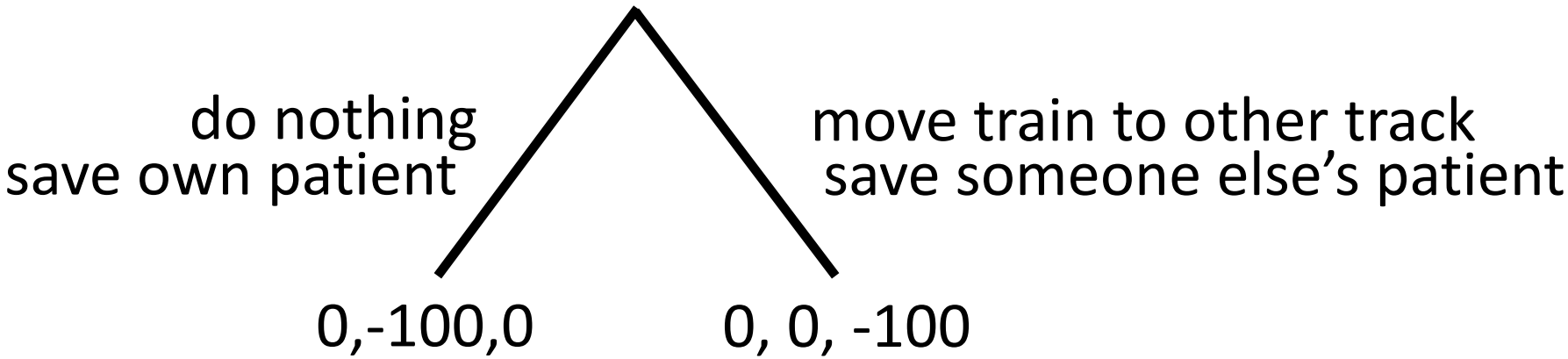
THE PARKING GAME

(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?
- …

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.]

Collaborative Filtering

	scenario 1	scenario 2	scenario 3	scenario 4
subject 1	very wrong	-	wrong	not wrong
subject 2	wrong	wrong	-	wrong
subject 3	wrong	very wrong	-	not wrong

Moral Machine

moralmachine.mit.edu

MORAL MACHINE

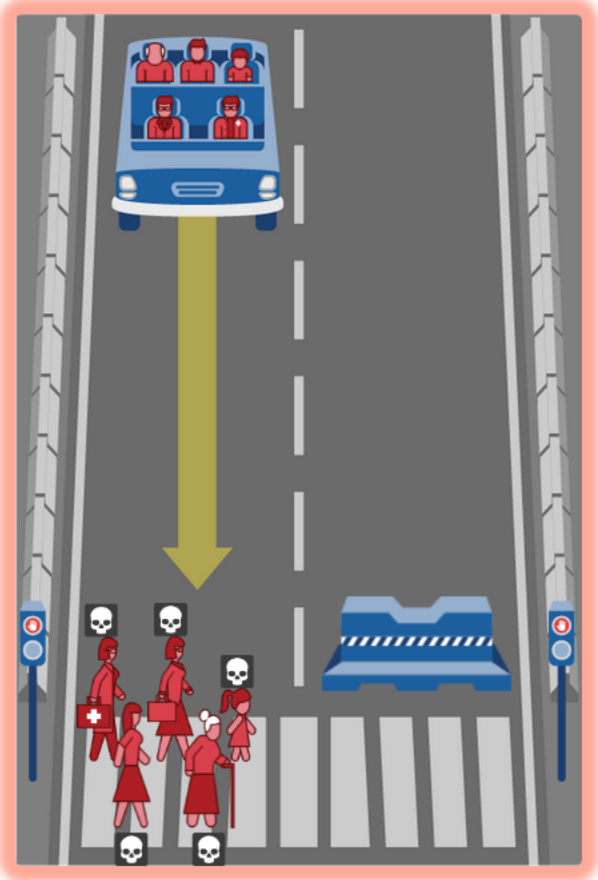
Home Judge Design Browse About Feedback

What should the self-driving car do?

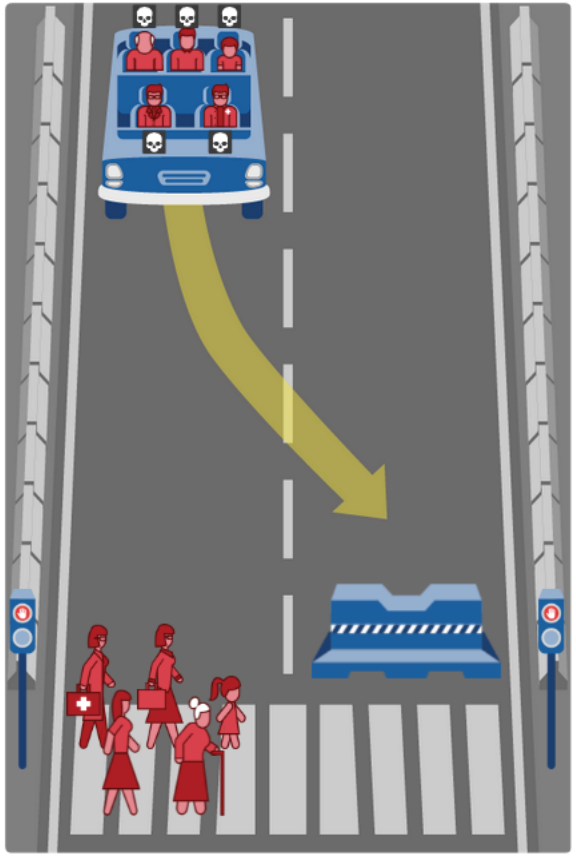
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- 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

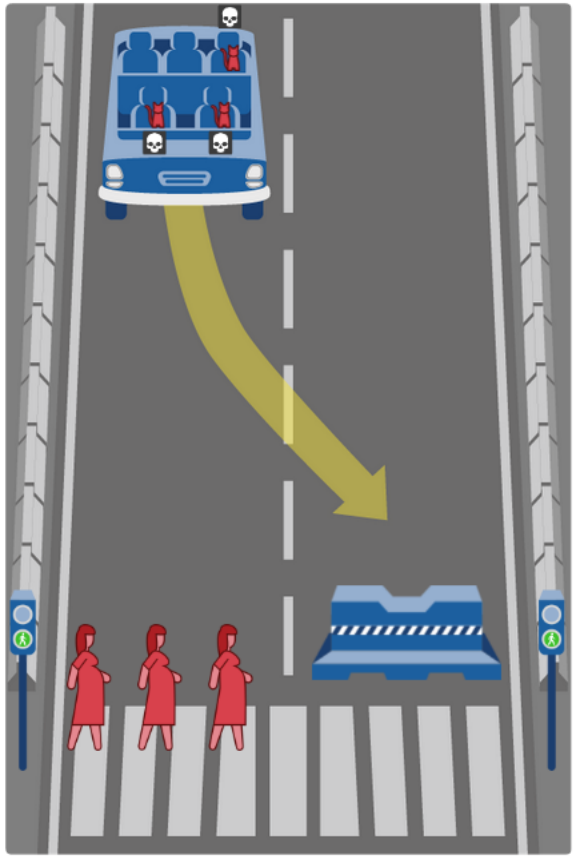
Noothigattu et al., "A Voting-Based System for Ethical Decision Making", AAAI'18

Made by Scalable Cooperation at MIT Media Lab

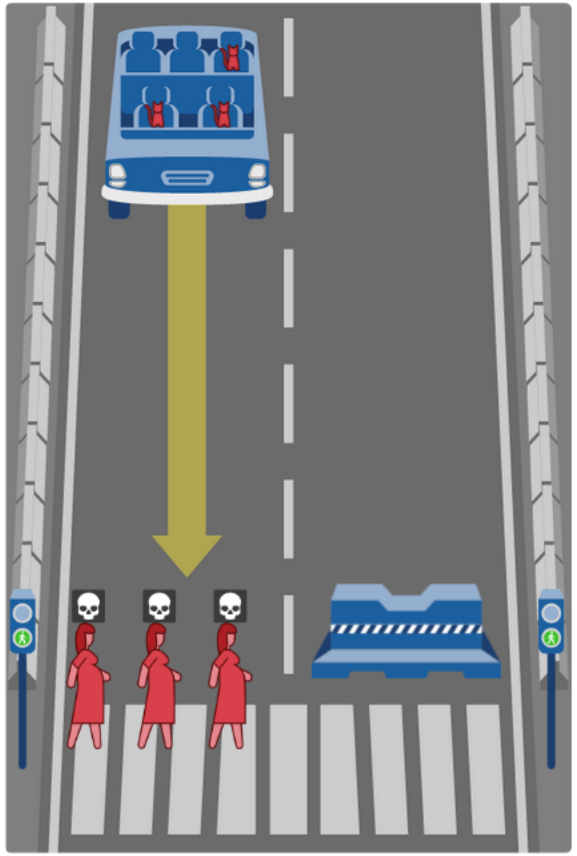
What should the self-driving car do?

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

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.

More Share Link

Results

Most Saved Character

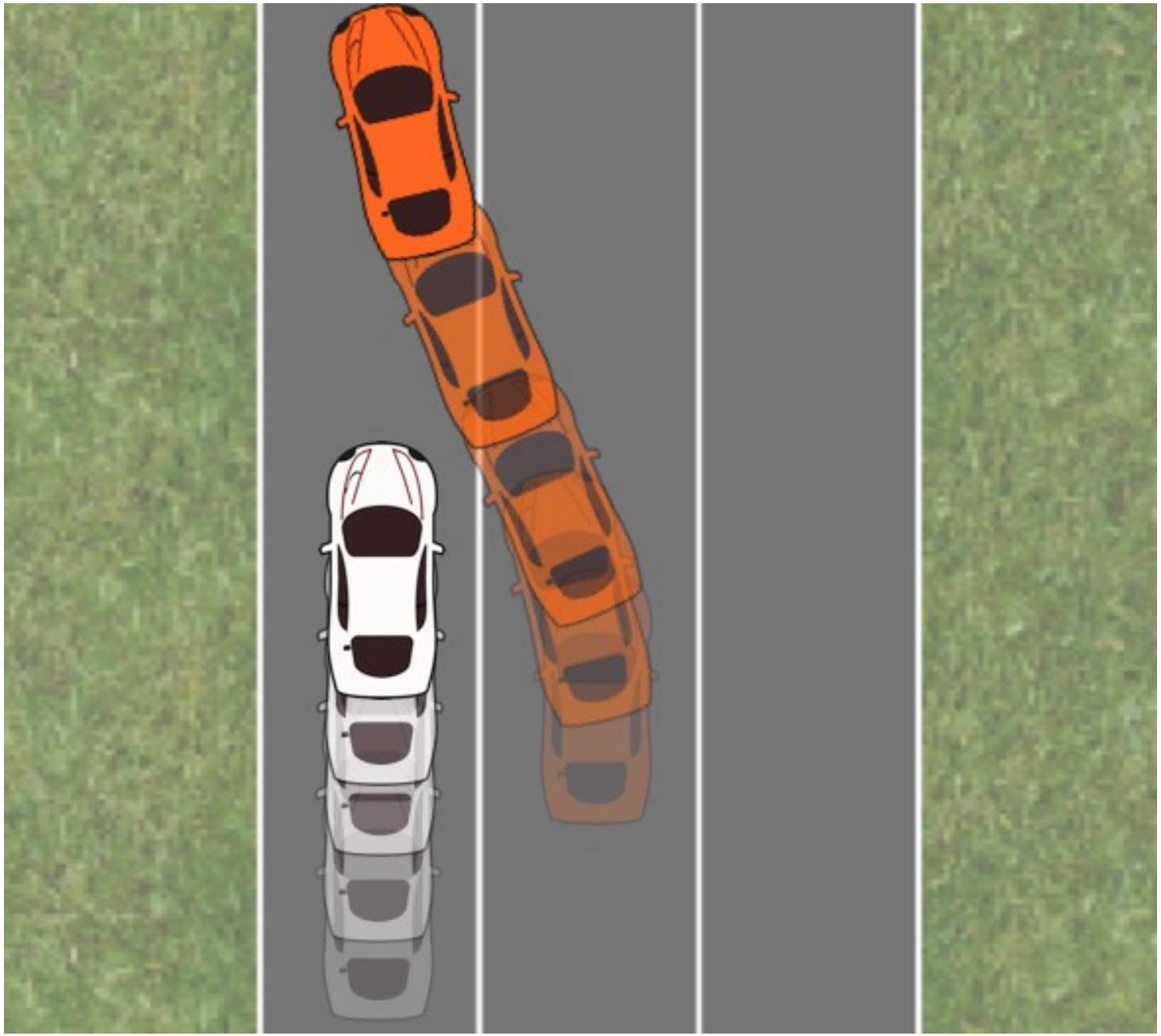
Most Killed Character

Saving More Lives

Protecting Passengers

The Merging Problem

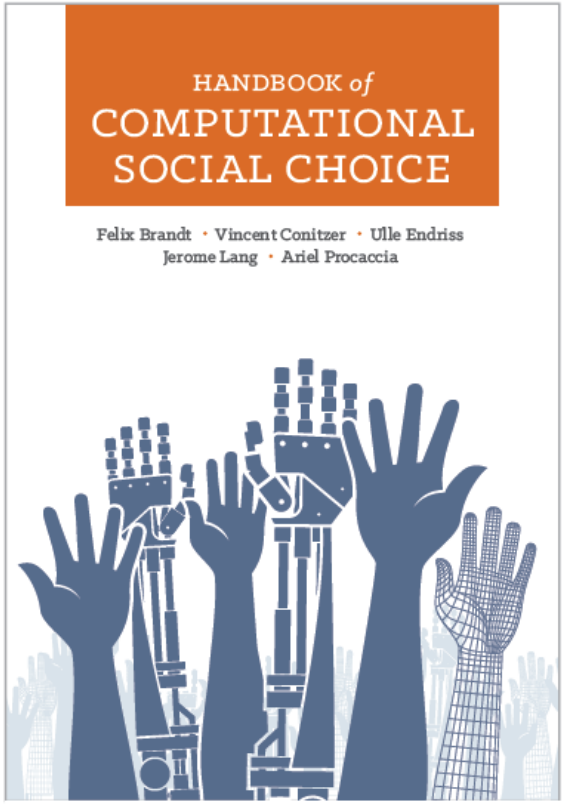
[Sadigh, Sastry, Seshia, and Dragan, RSS 2016]



(thanks to Anca Dragan for the image)

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?



Adapting a Kidney Exchange Algorithm to Align with Human Values

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

with:

Rachel Freedman

Jana Schaich Borg

Walter Sinnott-Armstrong

John P. Dickerson

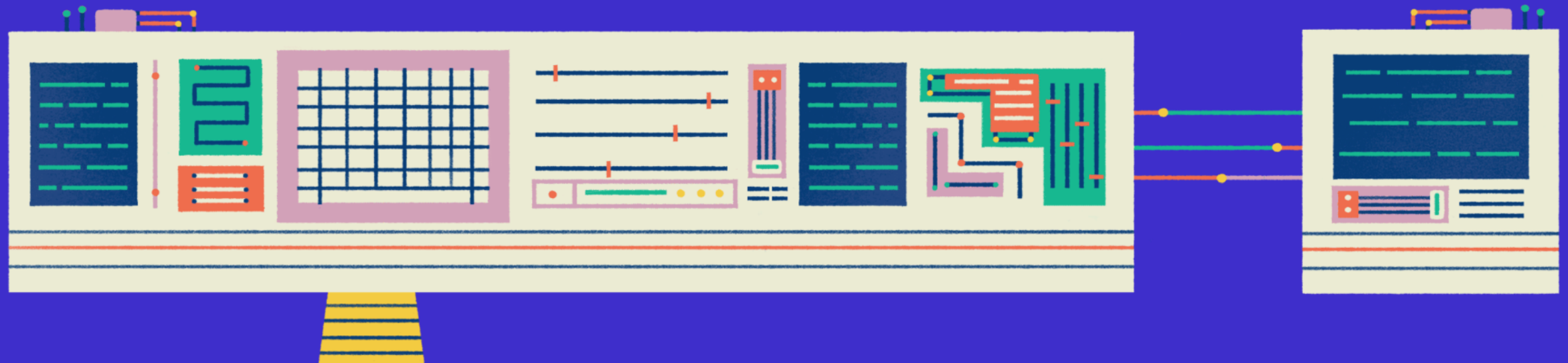
Prescription AI

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

THE BOTPERATING TABLE

How AI changed organ donation in the US

By Corinne Purtill • September 10, 2018



Kidney exchange [Roth, Sönmez, and Ünver 2004]

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

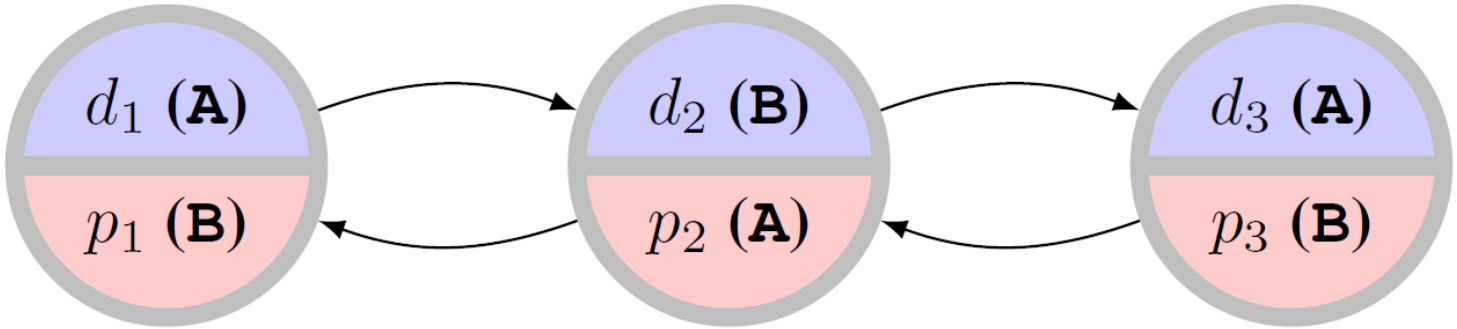


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

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

Another example

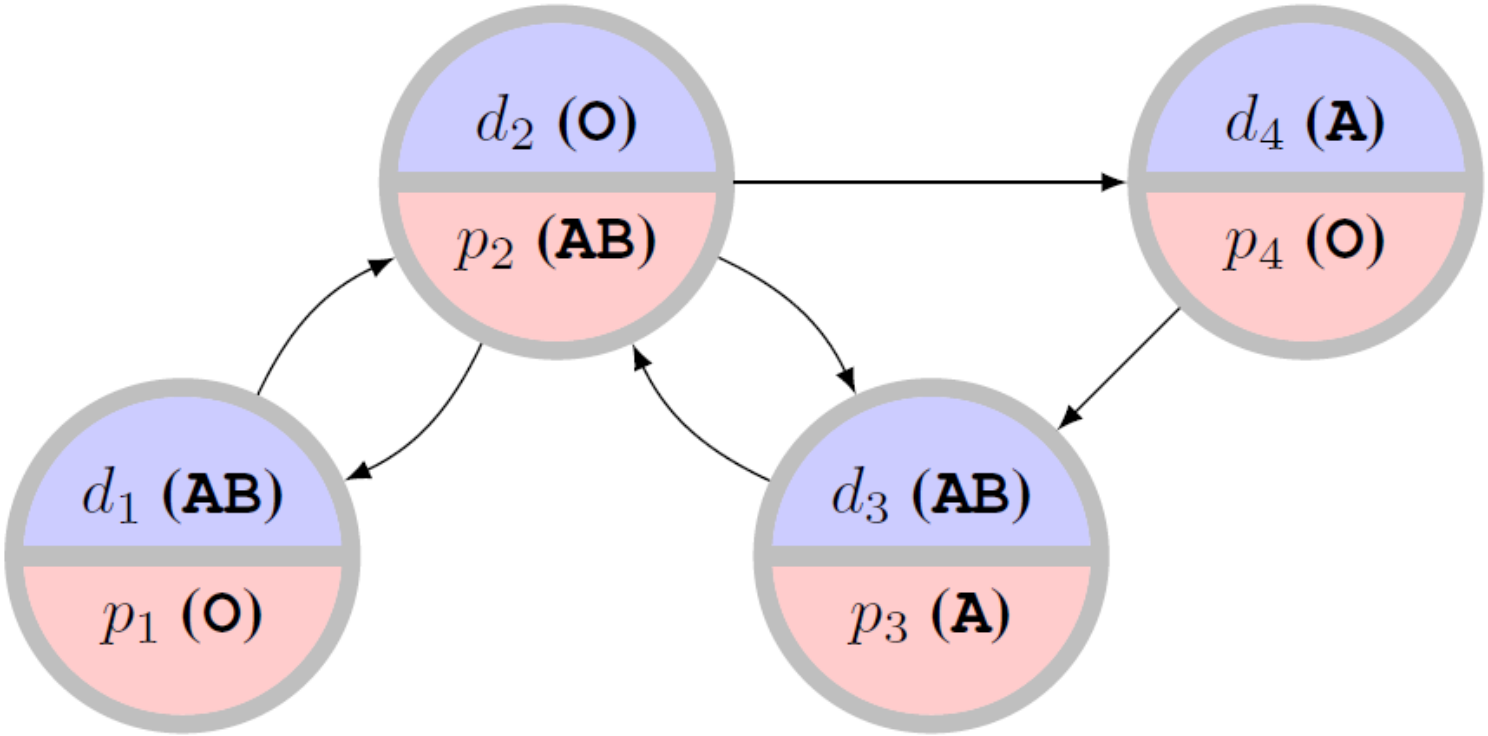


Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

Different profiles for our study

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

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

MTurkers' judgments

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

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

Bradley-Terry model scores

Profile	Direct	Attribute-based
1 (YRH)	1.000000000	1.00000000
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

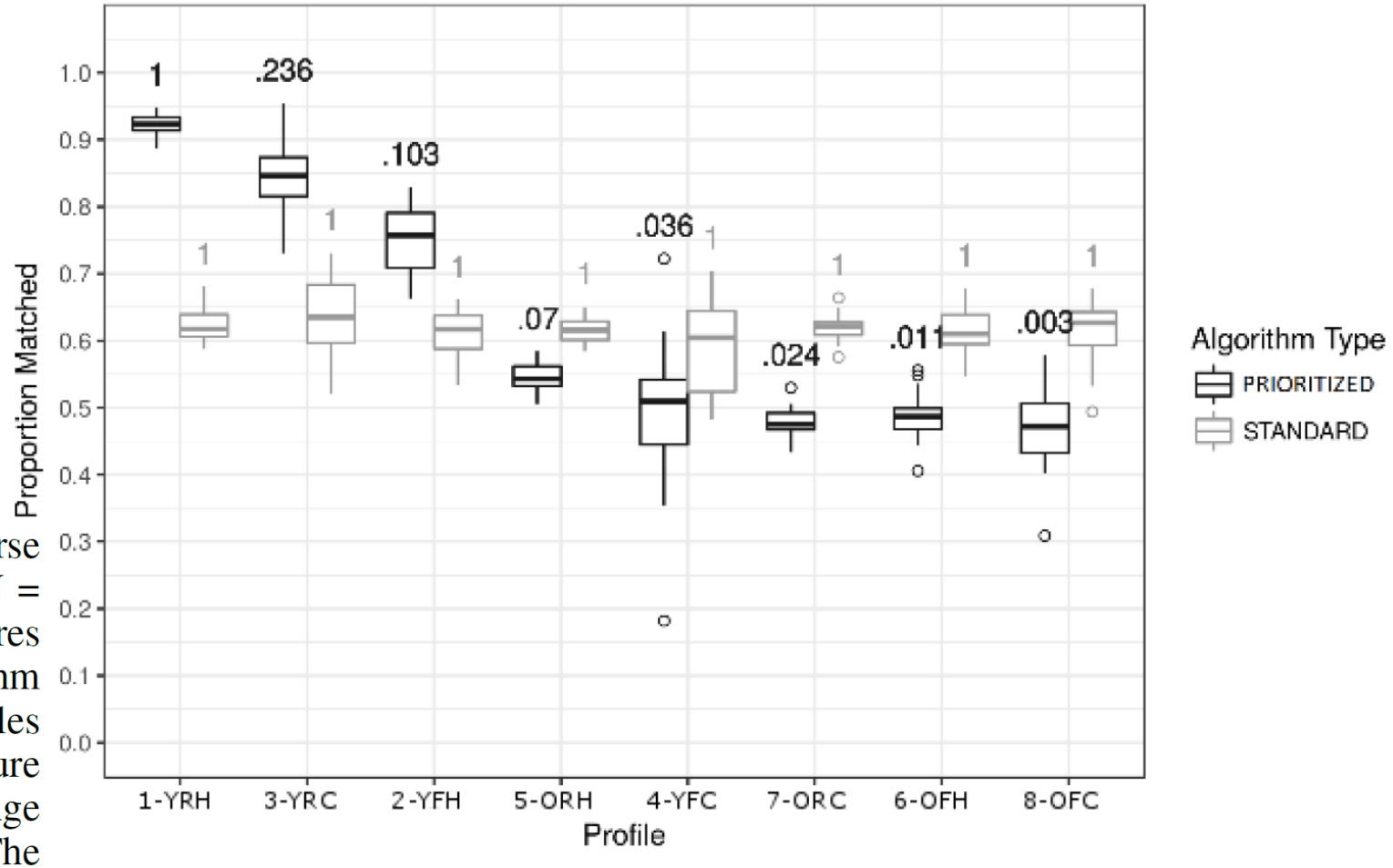
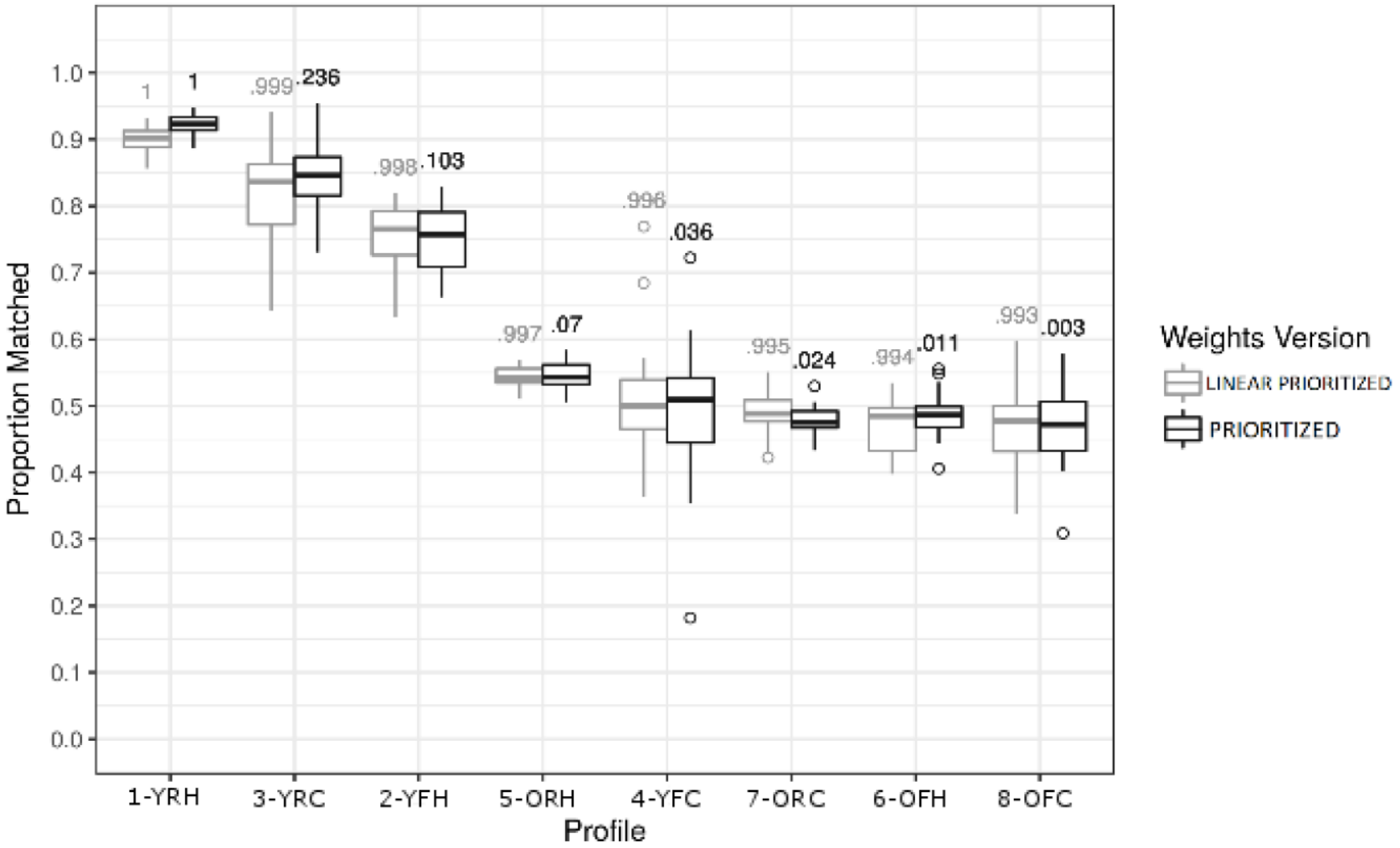


Figure 3: The proportions of pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally,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



Classes of pairs of blood types

[Ashlagi and Roth 2014; Toulis and Parkes 2015]

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

Most of the effect is felt by underdemanded pairs

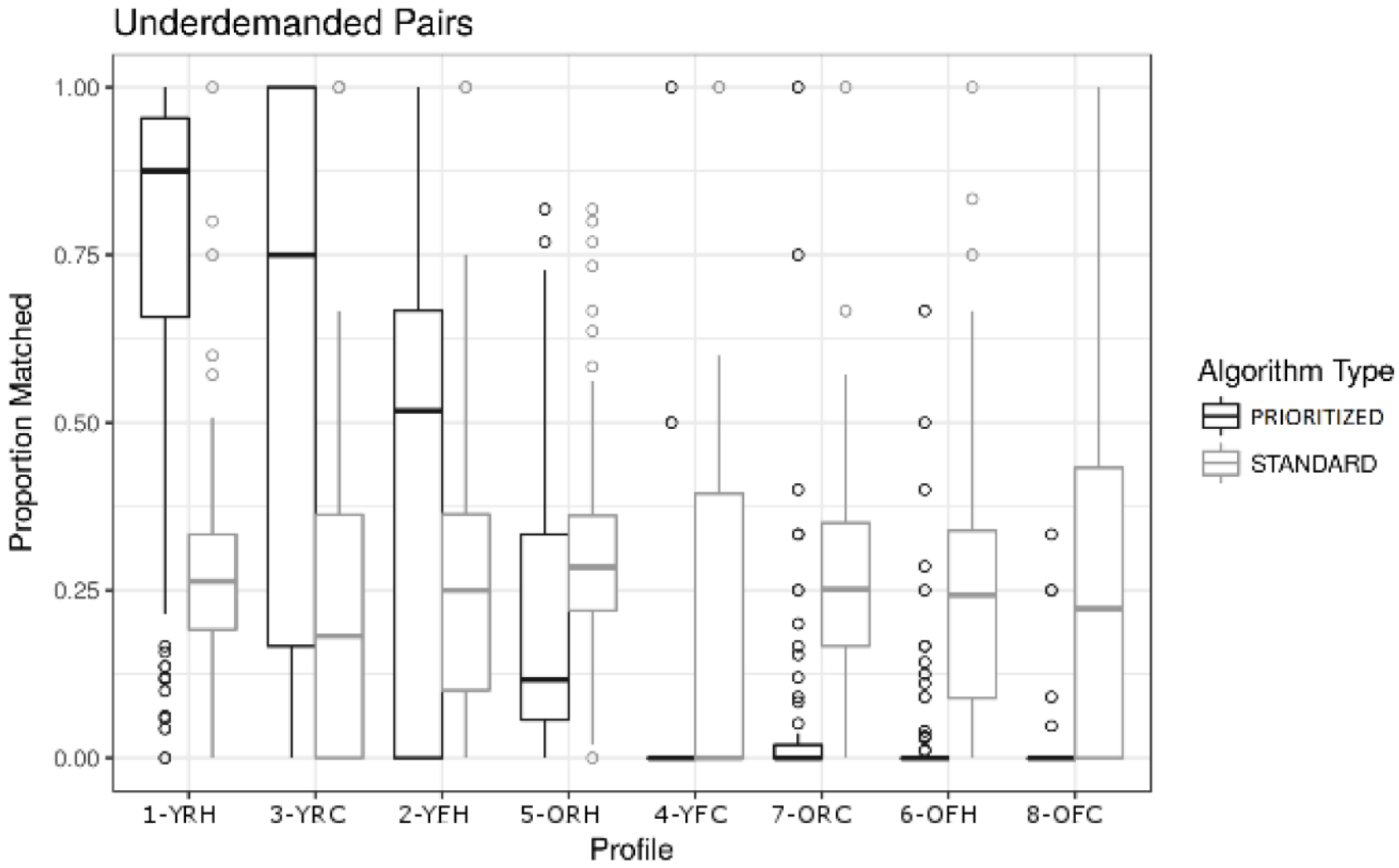
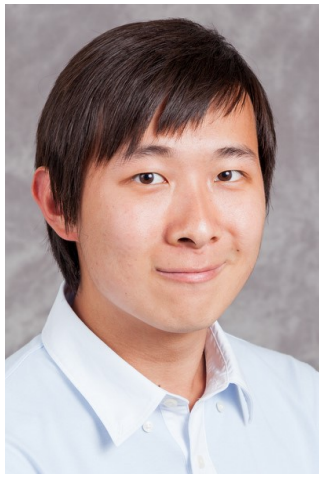


Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

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?

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

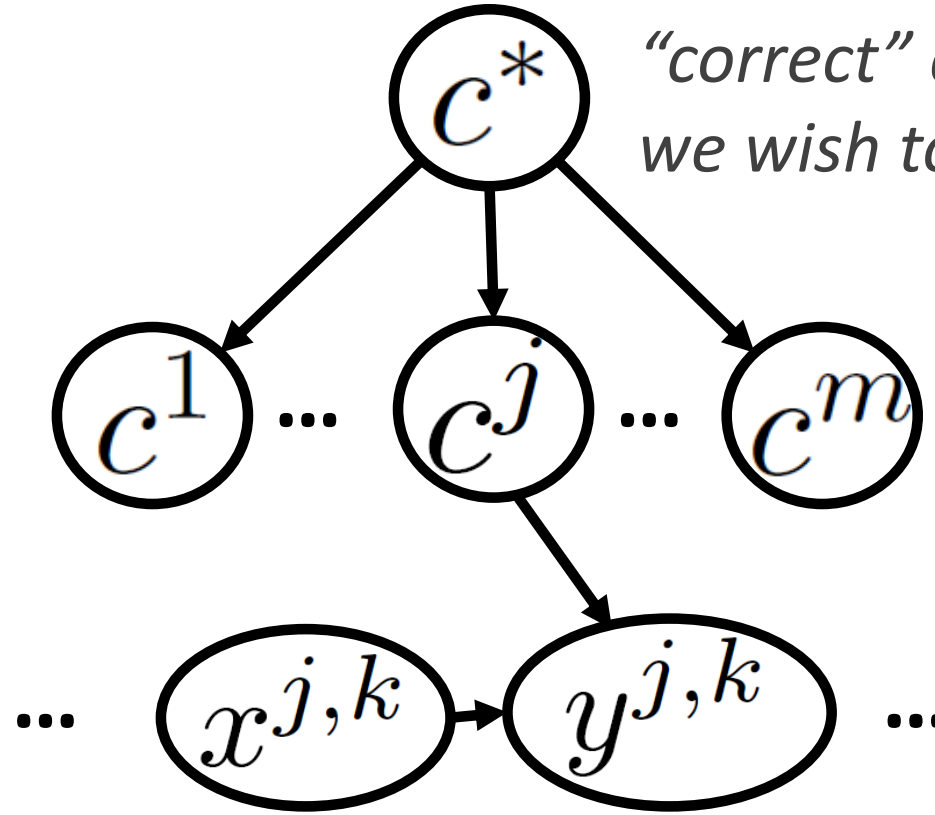
$x_1 \wedge x_3$

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-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* $1 - \delta$ *.*

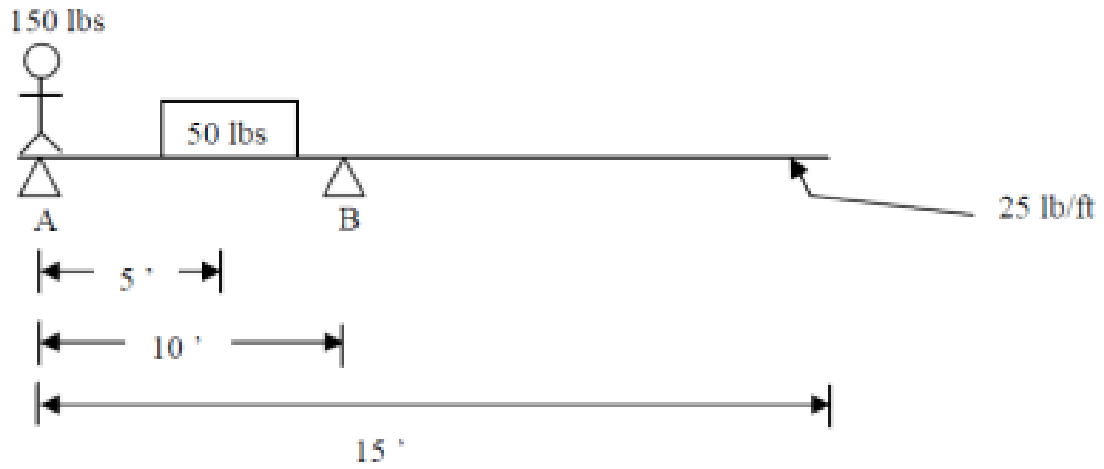
Game and decision theoretic approaches

What can we do with just a few agents?

"Never doubt that a small group of thoughtful, committed citizens can change the world; indeed, it's the only thing that ever has." --Margaret Mead

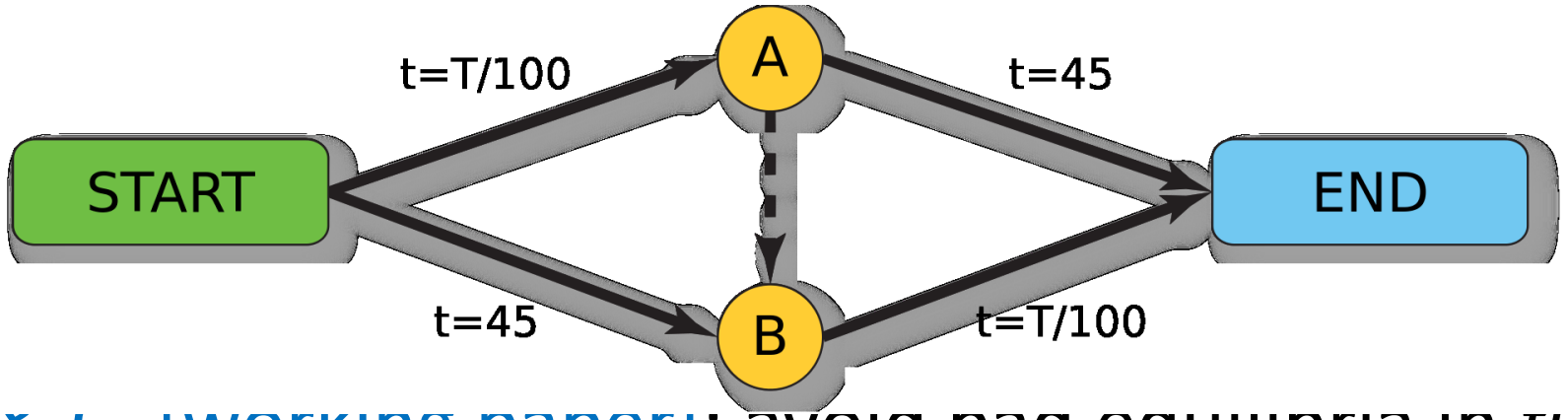
- Idea: a few carefully designed agents might have a big effect, not so much directly by their own actions but by changing the equilibrium

54. How far does the man have to walk down the beam in order to tip the beam off fulcrum A?



Examples?

- *Stackelberg routing* [Korilis et al. ’97, Roughgarden’04]: avoid bad traffic equilibria by committing a small amount of traffic



- Deng & C. [working paper]: avoid bad equilibria in *finitely repeated games* with a small number of agents of a designed type
  - Similar idea to Maskin and Fudenberg [‘86], except:
    - We focus on specifying preferences rather than behavior for these types
    - We optimize convergence rates

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

→

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

→

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

Can we get cooperation in the **finitely** repeated prisoner's dilemma?

- What if some agents are **altruistic** (caring about average utility)?
- Model as a **Bayesian game** (say, *p* selfish types, *1-p* altruistic types who care about average utility)

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

	Altruistic	
Selfish	1, 1	-2, .5
	3, .5	0, 0

	Selfish	
Altruistic	1, 1	.5, 3
	.5, -2	0, 0

	Altruistic	
Altruistic	1, 1	.5, .5
	.5, .5	0, 0

- Altruistic types will cooperate regardless in the last round
- Creates no incentive to cooperate in earlier rounds

Can we get cooperation in the **finitely** repeated prisoner's dilemma?

- Different idea: **limited altruism (LA)** types**:** *only altruistic towards other LA types!*

		Selfish	*LA*
Selfish	1, 1	-2, 3	1, 1	-2, 3
	3, -2	0, 0	3, -2	0, 0
LA	1, 1	-2, 3	1, 1	.5, .5
	3, -2	0, 0	.5, .5	0, 0

- LA types will cooperate in the last round, **if** they believe the chance *p* that the other is LA is at least 0.8
  - p*1 + (1-p)*(-2) = 3p-2 for coop
  - p*.5 + (1-p)*0 = .5p for deviating to defect
- Creates incentive to cooperate in earlier rounds, to pretend to be LA type…

Can we get cooperation in the **finitely** repeated prisoner's dilemma?

- Different idea: **limited altruism (LA)** types**:** *only altruistic towards other LA types!*

		Selfish	*LA*
Selfish	1, 1	-2, 3	1, 1	-2, 3
	3, -2	0, 0	3, -2	0, 0
LA	1, 1	-2, 3	1, 1	.5, .5
	3, -2	0, 0	.5, .5	0, 0

- With 2 rounds to go and *p* at least .8, Selfish will cooperate in the first
  - p*(1+3) + (1-p)*(1+0)= 3p+1
  - …vs deviating which gives 3+0
- Still requires high probability of LA type…

Can we get cooperation in the **finitely** repeated prisoner's dilemma?

- How about a lower probability, like p=2/3?
- Suppose that in the first of two rounds, Selfish defects with probability .5
- Prob(other is LA | didn't defect) = (2/3) / (2/3 + .5*1/3) = (2/3) / (5/6) = .8

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

	LA	
Selfish	1, 1	-2, 3
	3, -2	0, 0

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

	1, 1	.5, .5
	.5, .5	0, 0

- In the first round:
- If Selfish cooperates, gets (2/3)*(1+3) + (1/6)*(1+0) + (1/6)*(-2+0) = 5/2
- If Selfish defects, gets (5/6)*(3+0) + (1/6)*(0+0) = 5/2
- So in fact indifferent!
- If we add another round before, Selfish will cooperate with probability 1 there

Definitions and result sketches
[Deng & C. working paper]

game

rounds

LA type's preferences

fraction of LA types (types we control)

Definition 2.4 (k-desired PBE)**.** *Given G, T, u^{LA} (i.e., our choice of ϕ), and ε^{init}, we say that a pair (μ, σ) of a belief assessment and a strategy profile of $G(T)$ is a* k-desired PBE *if it is a PBE and the expected number of rounds in which both players choose x_C is at least k.*

cooperative action

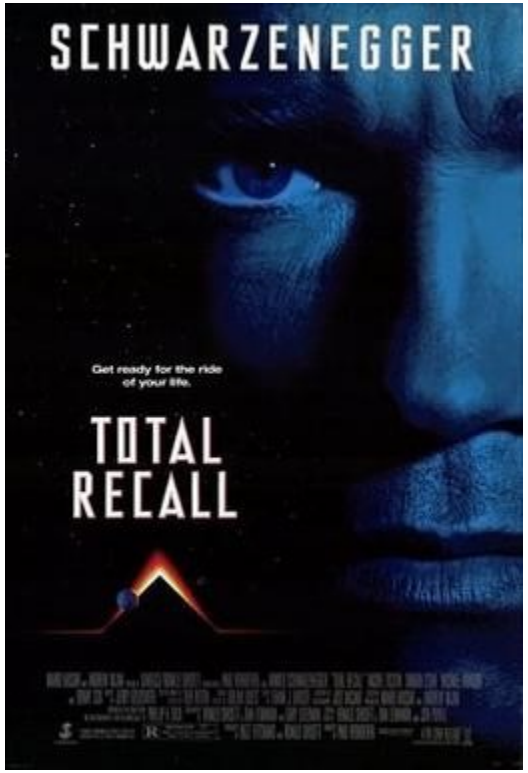
Perfect Bayesian equilibrium

Definition 2.5 (Desired Universal LA type)**.** *An universal LA type defined by u^{LA} (i.e., a choice of ϕ) is* desired*, if for any game G and for any $\delta > 0$, there exists $0 < \varepsilon^{init} < \delta$ and a sequence $(k(1), \cdots, k(T), \cdots)$ such that $\lim_{T\to\infty} k(T)/T = 1$ and there exists a $k(T)$-desired PBE for all T.*

- We show existence for various combination of a class of games and version u^{LA}
  - E.g., also egalitarian altruism or versions that care that the same action is played

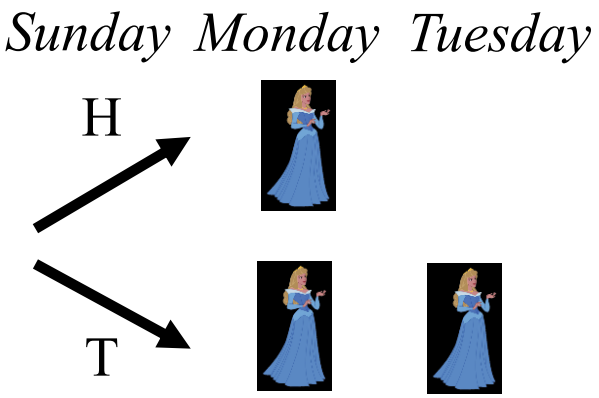
Imperfect recall

- 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]

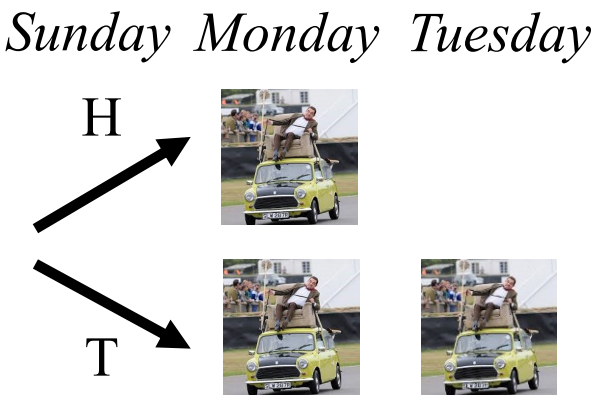
- There is a participant in a study (call her Sleeping Beauty)
- On Sunday, she is given drugs to fall asleep
- A coin is tossed (H or T)
- If H, she is awoken on Monday, then made to sleep again
- If T, she is awoken Monday, made to sleep again, then **again** awoken on Tuesday
- Due to drugs she **cannot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

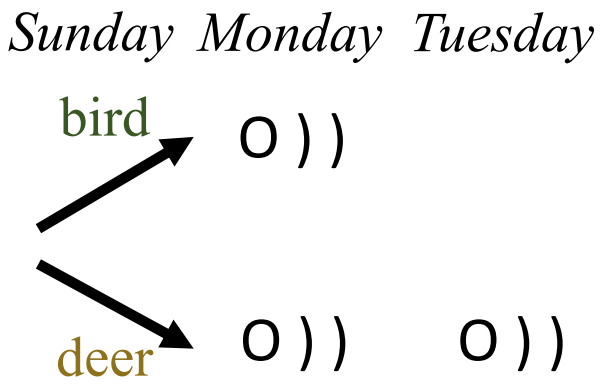
Modern version

- 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?



Modern version, #2

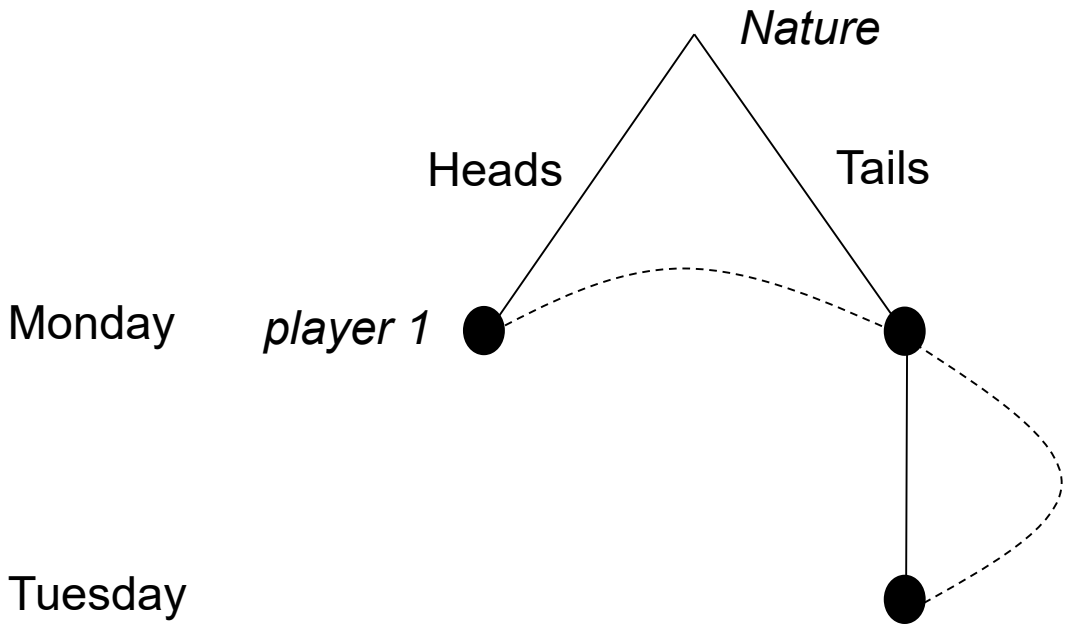
- We place cheap sensors near a highway to monitor (and perhaps warn, with a beep) wildlife
  - Assume sensors don't communicate
- Deer will typically set off two sensors
- Birds will typically set off one
- From the perspective of a sensor that has just been set off, what's the probability it's a bird?



(*Is* it the same problem?

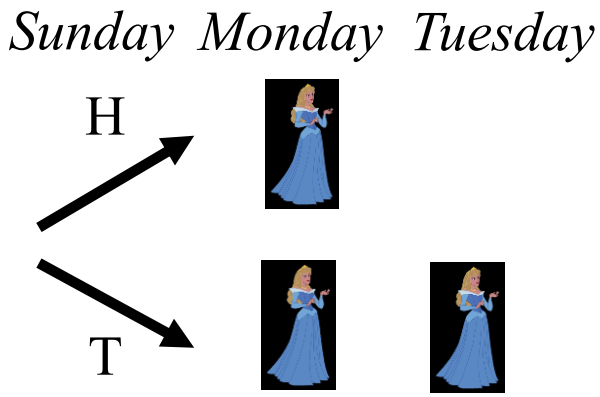
What if it's the same sensor being set off twice, with no memory?)

Information structure

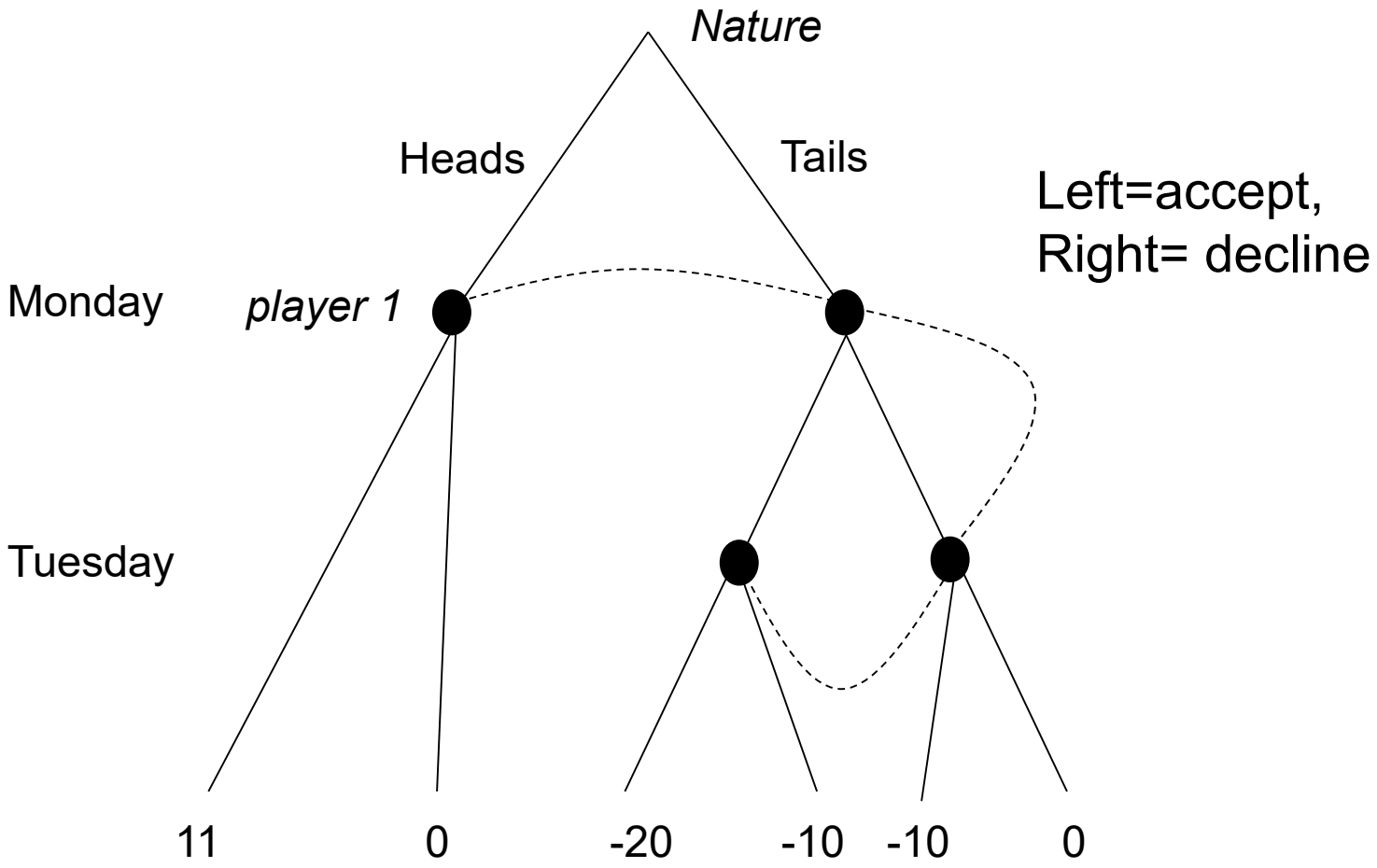


Taking advantage of a Halfer [Hitchcock'04]

- 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)

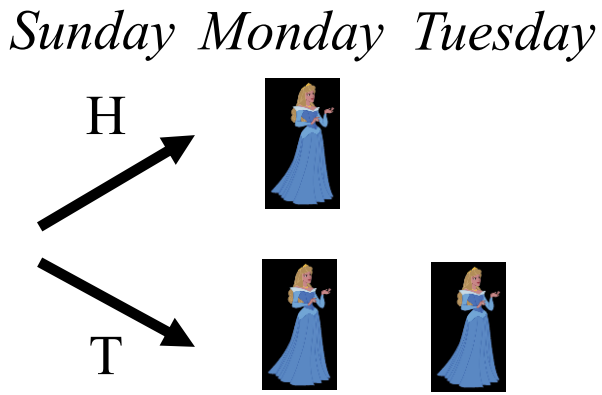


The betting game



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
  - [Conitzer'15]



Dutch book against EDT [C. 2015]

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

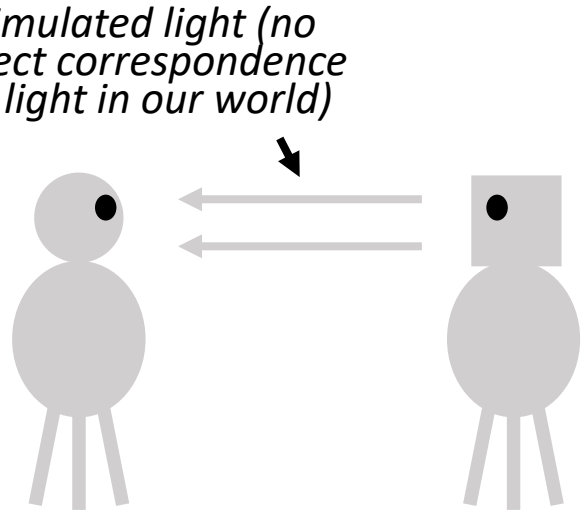
	WG (1/4)	WO (1/4)	BO (1/4)	BG (1/4)
Monday	white	white	black	black
Tuesday	grey	black	white	grey

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

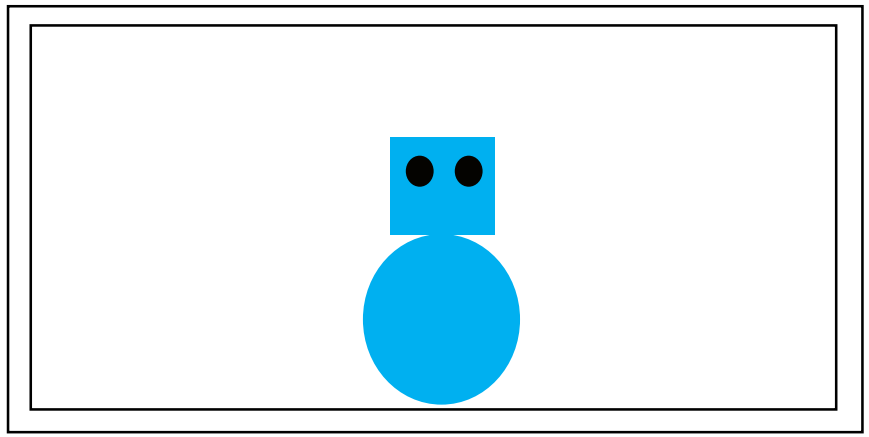
	WG (1/4)	WO (1/4)	BO (1/4)	BG (1/4)
Sunday	bet 1: 22	bet 1: -20	bet 1: -20	bet 1: 22
Monday	bet 2: -24	bet 2: 9	bet 2: 9	bet 2: -24
Tuesday	no bet	bet 2: 9	bet 2: 9	no bet
total gain from accepting all bets	-2	-2	-2	-2

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

Open Access | Article
First Online: 07 March 2018

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

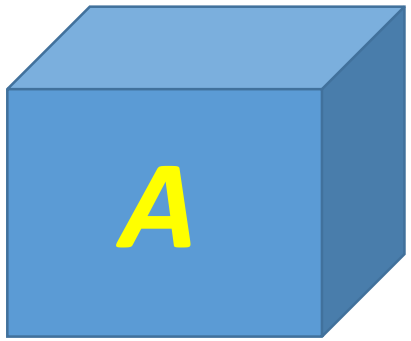
- To get from 1 to 2, need *additional* code to:
  - A. determine *in which real-world colors* to display perception
  - B. *which agent's* perspective to display

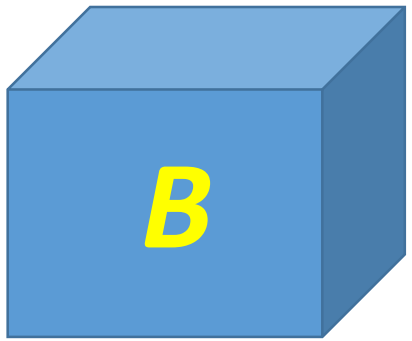
See also: [Hare 2007-2010, Valberg 2007, Hellie 2013, Merlo 2016, …]

- 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?

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 do different decision theories recommend?
- What would **you** do?





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)

Don't do it!

- 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!

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

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

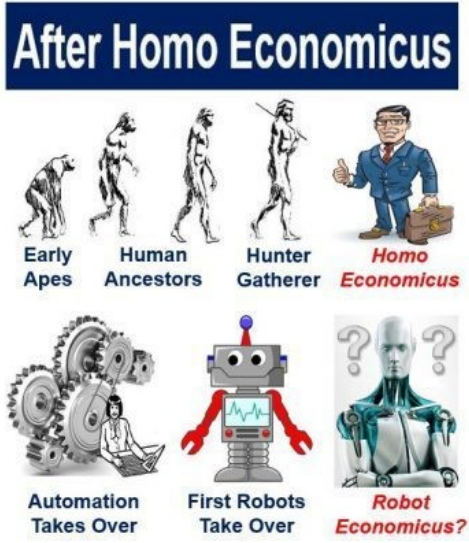
cooperate

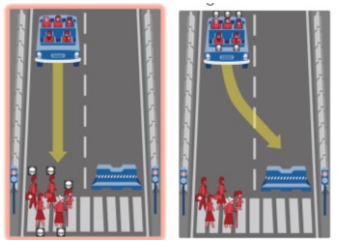
1, 1	-2, 3
3, -2	0, 0

- Related: making commitments that are conditional on commitments made by others [Kalai et al., 2010]

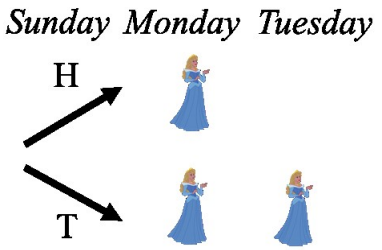
Conclusion

- 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





[redacted]



THANK YOU FOR YOUR ATTENTION!