

New Directions in Belief Formation and Decision Theory for AI

Theory for AI

Vincent Conitzer
Duke University

Various parts are joint work with:



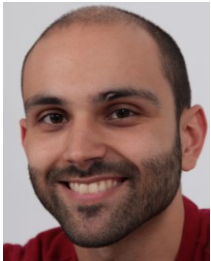
Caspar Oesterheld



Scott Emmons



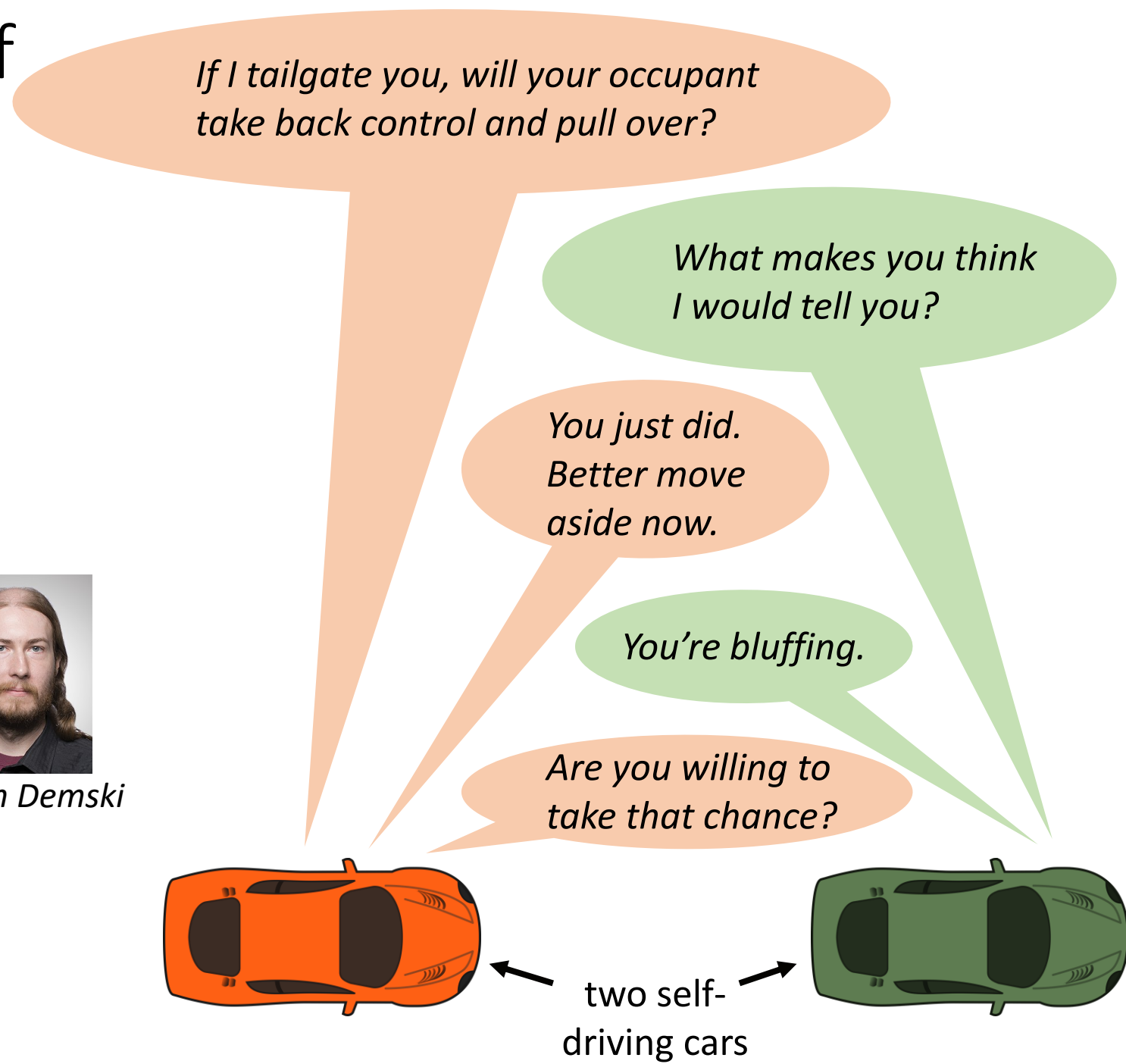
Abram Demski



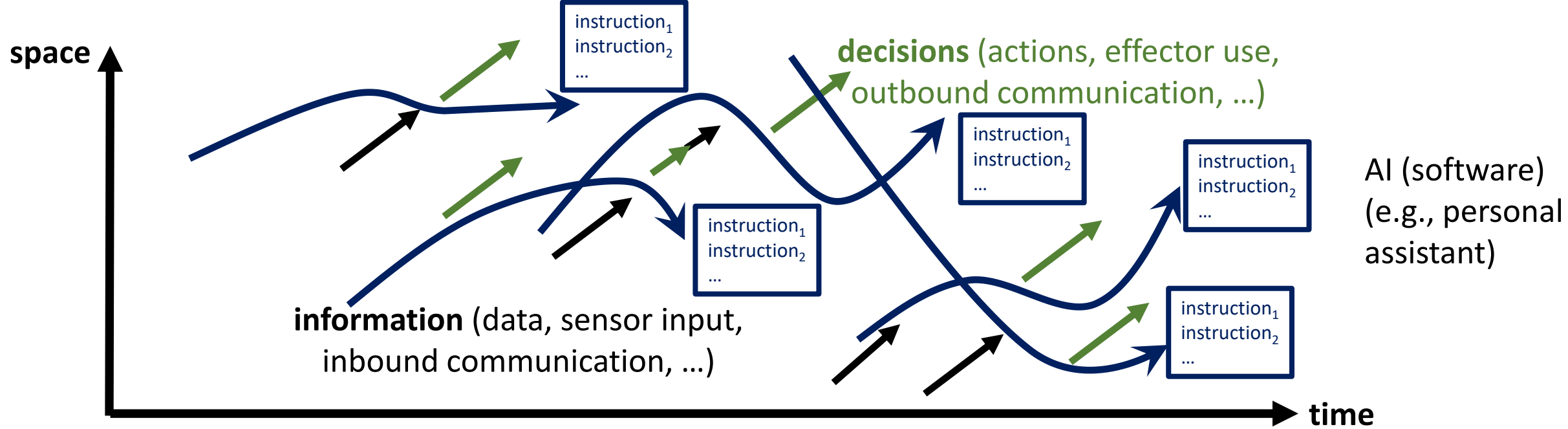
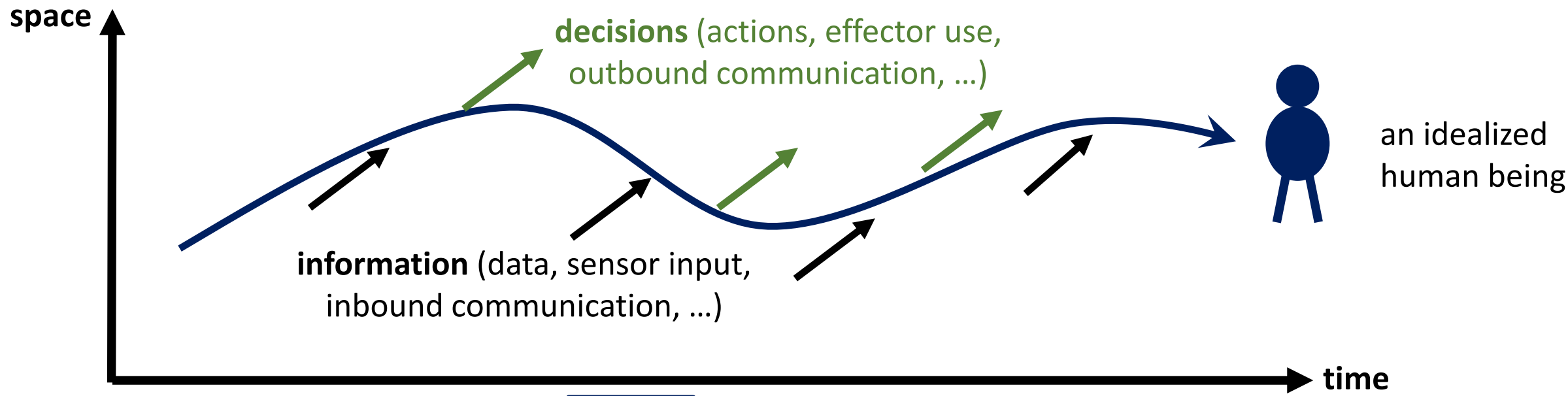
Andrew Critch



Stuart Russell

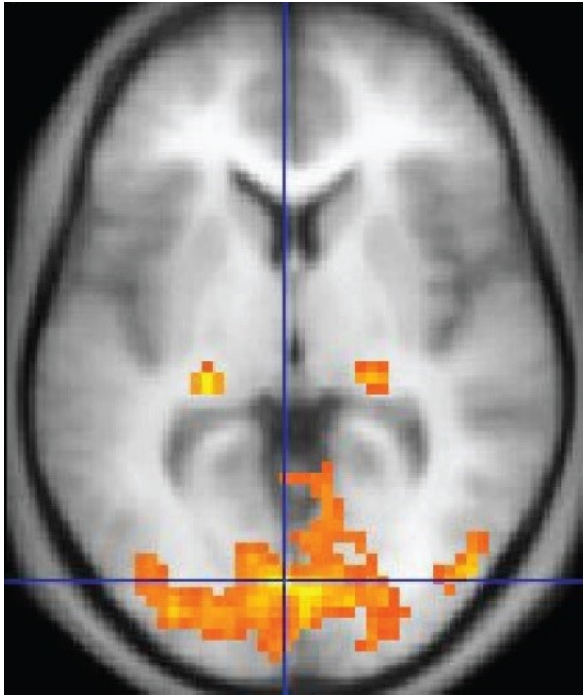


Agents through time



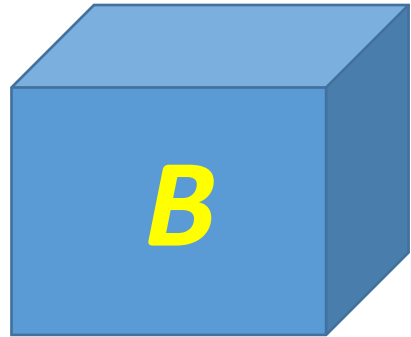
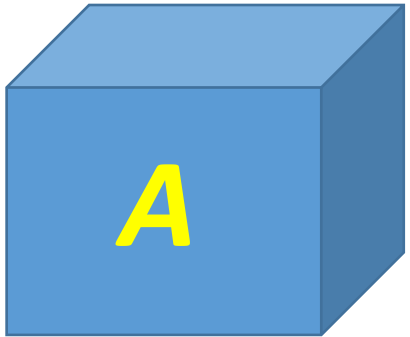
What should you do if...

- ... 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?

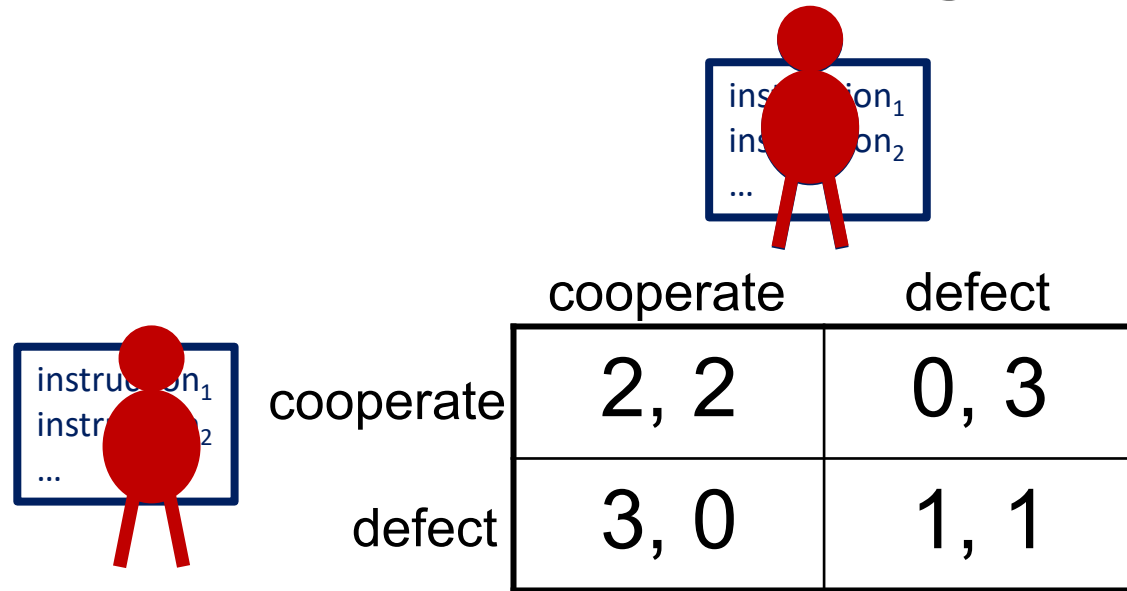


related to working paper
[\[Oesterheld and C.\]](#)



Caspar Oesterheld

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**...

- ... for what the demon put in the boxes
- ... for whether your twin defects
- ... for whether you eat the cookie on the other day



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



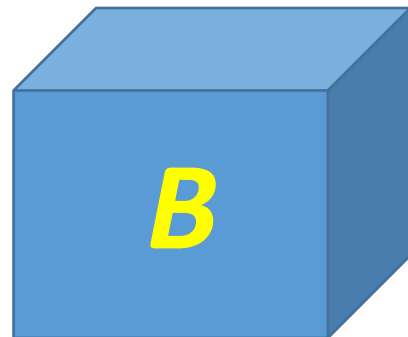
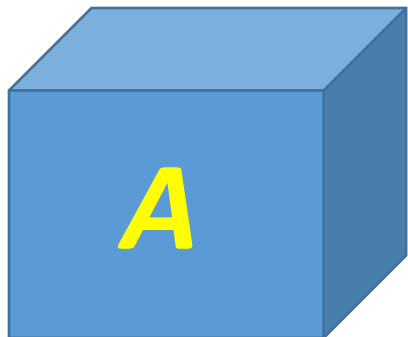
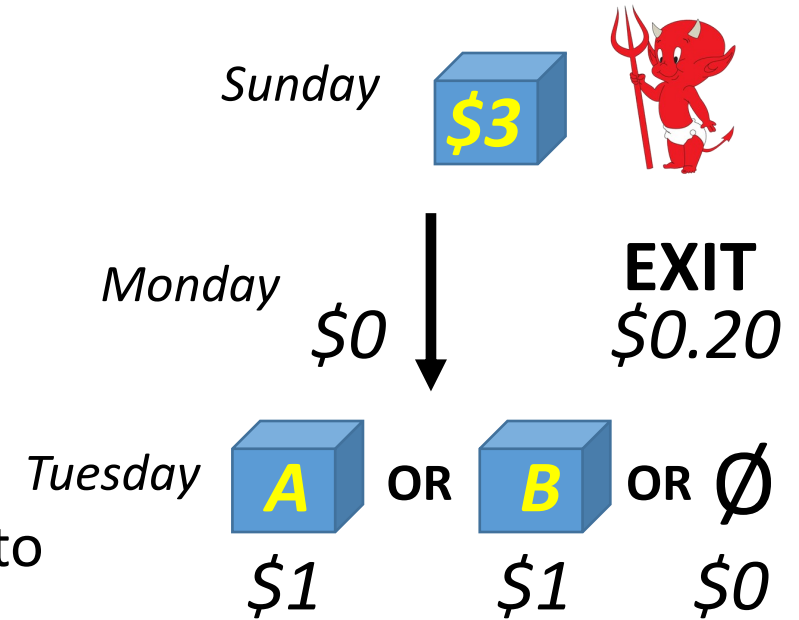
- *Evidential Decision Theory (EDT)*: When considering how to make a decision, consider **how happy you expect to be conditional on taking each option** and choose an option that maximizes that
- *Causal Decision Theory (CDT)*: Your decision should focus on what you **causally affect**

Turning causal decision theorists into money pumps

[Oesterheld and C., working paper]



- **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





About MAA Membership MAA Publications Meetings Competitions Programs and Communities

- MAA Publications
- Periodicals
 - The American Mathematical Monthly
 - Mathematics Magazine
 - The College Mathematics Journal
 - Loci/JOMA
 - Convergence
 - MAA FOCUS

Home » MAA Publications » Periodicals » The American Mathematical Monthly » American Mathematical Monthly - August/September 2017

American Mathematical Monthly - August/September 2017



Enjoy the lazy days of summer and some engaging mathematics in the latest issue of the *Monthly*.

Peter Winkler explores the probabilistic and philosophical conundrums facing Sleeping Beauty and those observing her as she is awakened once or twice during her slumber. Arseniy Akopyan and Vladislav Vysotsky study the relation between the length of a curve that passes through a fixed number of points on

Quick Links

[Become a Member](#)

[Register for AM](#)

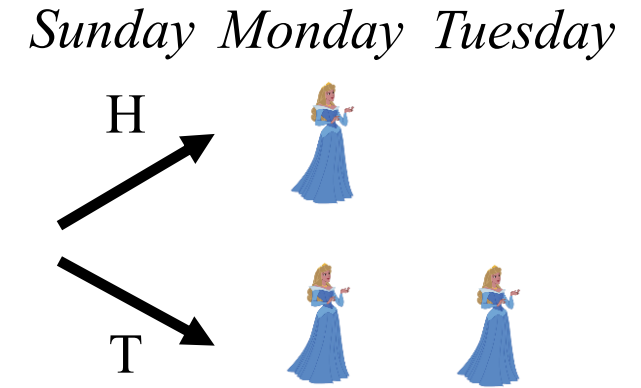
Member Publication

As a member of MAA you have access to premier publications like:

[The American Mathematical Monthly](#)

The Sleeping Beauty problem [Elga, 2000]

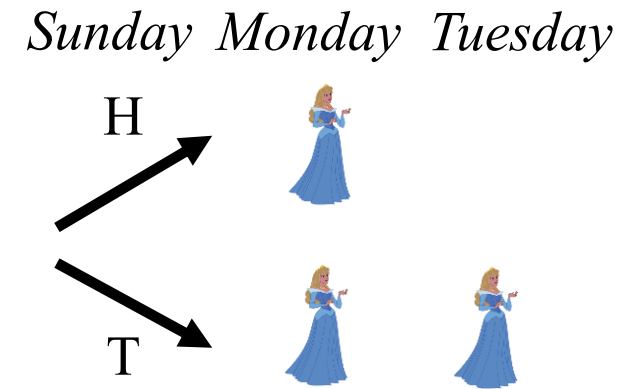
- There is a participant in a study (call her Sleeping Beauty)
- On Sunday, she is given drugs to fall asleep
- A coin is tossed (H or T)
- If H, she is awoken on Monday, then made to sleep again
- If T, she is awoken Monday, made to sleep again, then **again** awoken on Tuesday
- Due to drugs she **cannot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

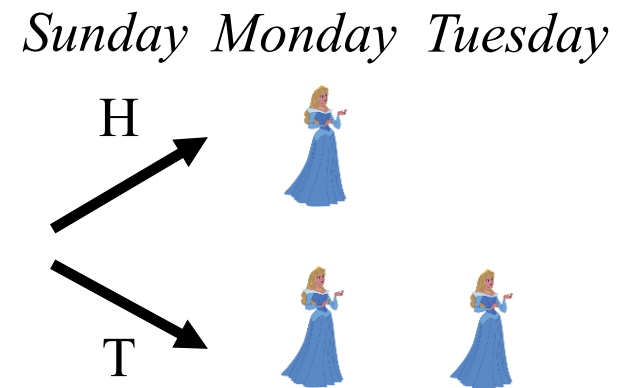
Taking advantage of a Halfer [\[Hitchcock'04\]](#)

- 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]



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

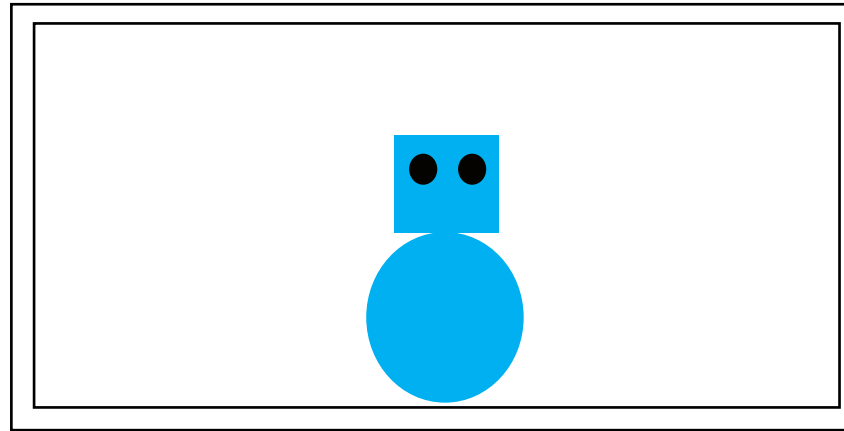
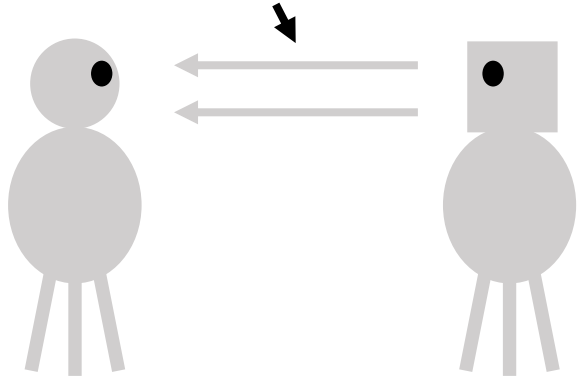
Fig. 3 Sequences of coin tosses and corresponding room colors, as well as their probabilities, in the WBG Sleeping Beauty variant.

	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

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

[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 Shares 3.7k Downloads 1 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](#), ...]

- 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?

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 ...

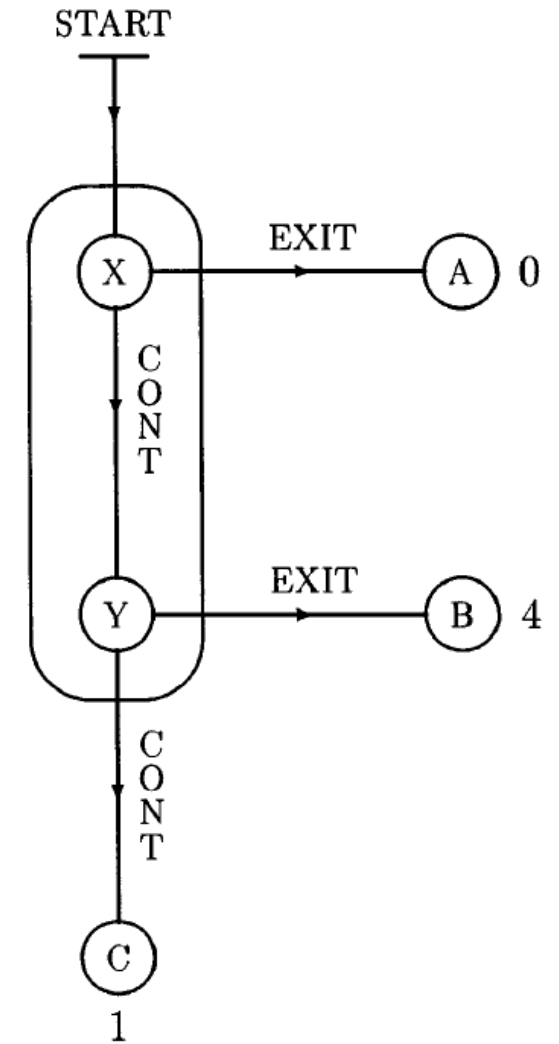


FIG. 1. The absent-minded driver problem.

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

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

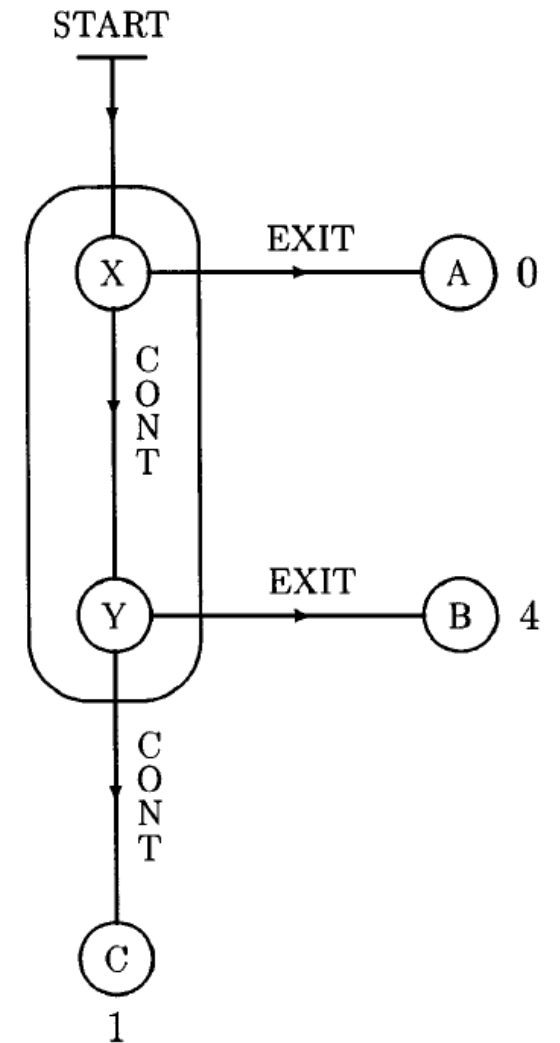
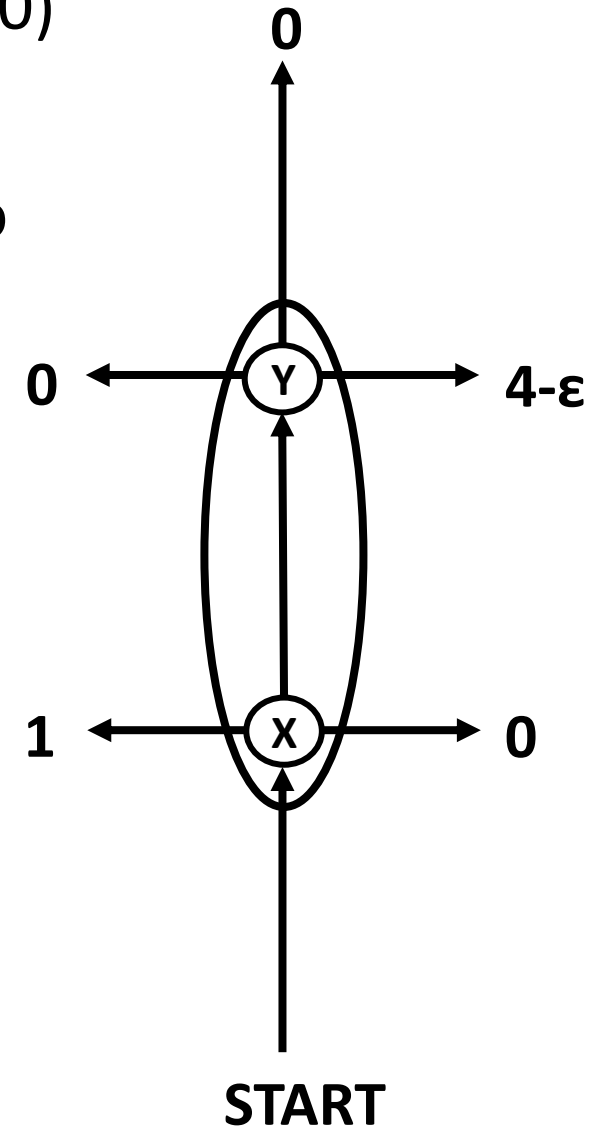


FIG. 1. The absent-minded driver problem.

Image from Aumann, Hart, Perry 1997

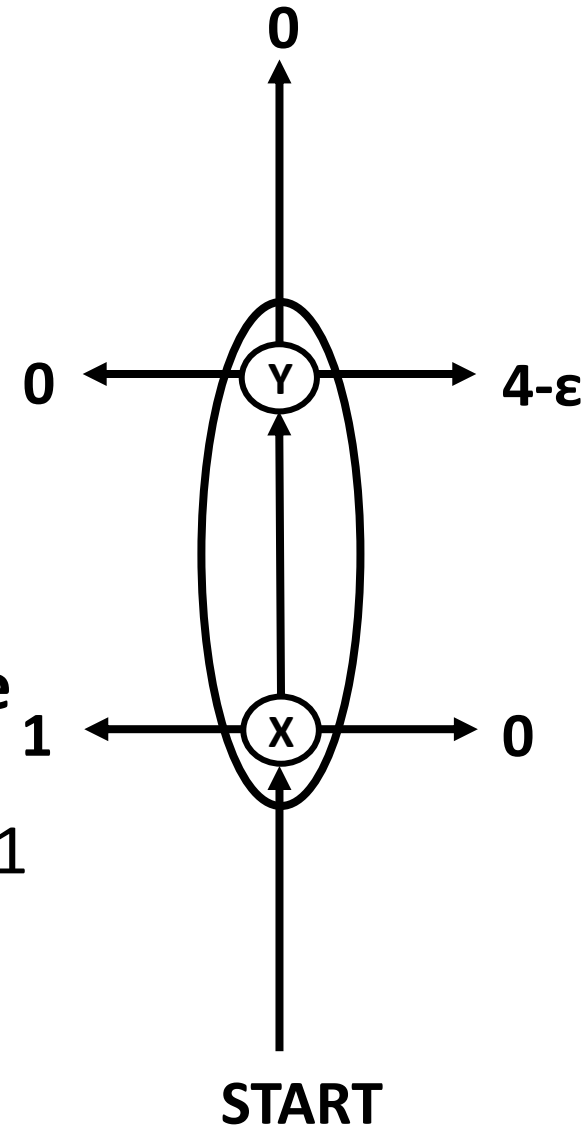
A challenging example for the evidential decision theorist

- 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 $(p_l, p_s, p_r) = (0, \frac{1}{2}, \frac{1}{2})$, 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: $b * \frac{1}{2} * (4 - \varepsilon) + (1 - b) * \frac{1}{4} * (4 - \varepsilon) = 1 + b - \frac{1}{4}\varepsilon - \frac{1}{4}b\varepsilon$
- For sufficiently small ε this is greater than 1
- Hence EDT suggests $(0, \frac{1}{2}, \frac{1}{2})$ 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 $(0, \frac{1}{2}, \frac{1}{2})$, in a halving sort of way
- $P(XY(4-\epsilon), @X) = P(XY(4-\epsilon)) * P(@X | XY(4-\epsilon)) = \frac{1}{4} * \frac{1}{2}$
- $P(XY(4-\epsilon), @Y) = P(XY(4-\epsilon)) * P(@Y | XY(4-\epsilon)) = \frac{1}{4} * \frac{1}{2}$
- Any other trajectory with positive probability gives payoff 0
- So expected utility is $2 * \frac{1}{4} * \frac{1}{2} * (4-\epsilon) = 1 - \epsilon/4$, 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
- $P(XY(4-\epsilon), @X) = \frac{1}{4} * \frac{1}{2}$; $P(XY(0), @X) = \frac{1}{4} * \frac{1}{2}$; $P(X(0), @X) = \frac{1}{2} * 1$
- So $P(X(0) | @X) = \frac{1}{2} / (\frac{1}{2} + \frac{1}{4}) = \frac{2}{3}$ (**not** $\frac{1}{2}$)
- Previous slide had **hidden assumption**: *where I am carries no information about my **future** coin tosses*



Functional Decision Theory

[Soares and LeVine 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
```

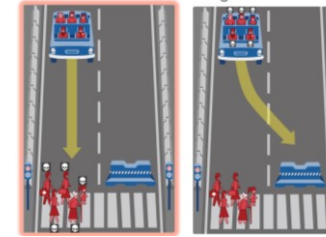
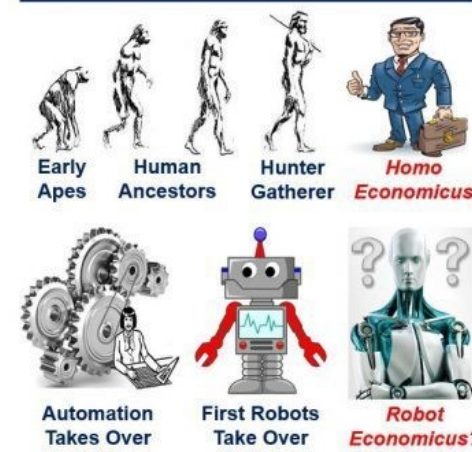


	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

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

Conclusion

After Homo Economicus



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

 \rightarrow

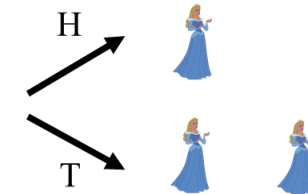
1, 1	-2, 3
3, -2	0, 0

 \rightarrow

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



Sunday Monday Tuesday



- 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
- \rightarrow 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!