15-784 (CMU)
Foundations of Cooperative AI

you should also be on Piazza for the course

See also: Foundations of Cooperative AI Lab (FOCAL)

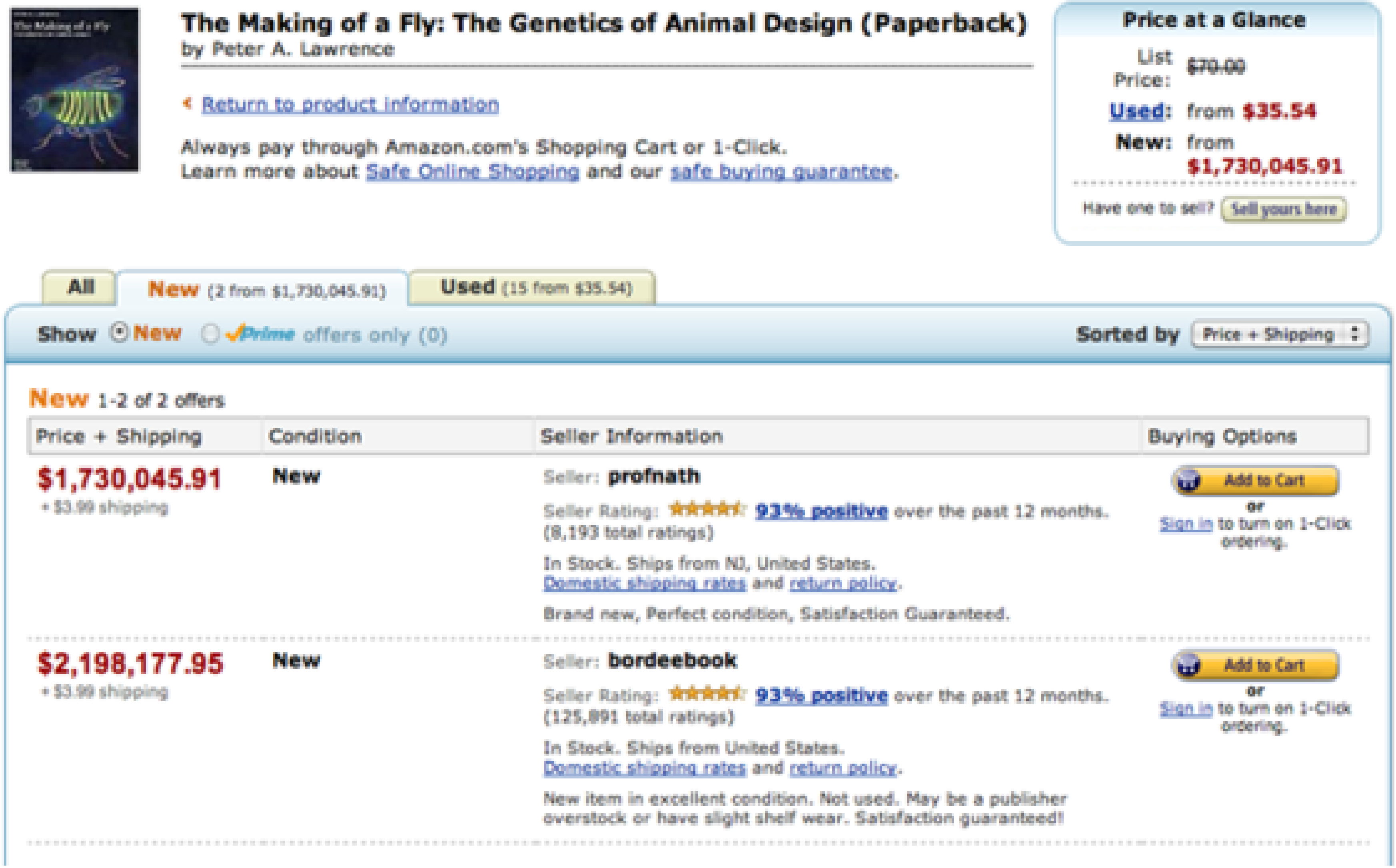
Vision paper: Vincent Conitzer and Caspar Oesterheld. Foundations of Cooperative AI. AAAI'23

Vincent Conitzer
(instructor)

Emanuel Tewolde
(TA)

Outline for today / this week

- Logistics
- High-level intro / motivation
- Quickly touching on all kinds of topics in the course!
  - Start identifying possible projects from day 1
  - Let us know if you want more references on any topic!



The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance

List Price: ~~$70.00~~
Used: from **$35.54**
New: from **$1,730,045.91**

Have one to sell? Sell yours here

All | **New** (2 from $1,730,045.91) | **Used** (15 from $35.54)

Show ⦿ New ○ ✓Prime offers only (0) — Sorted by Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$1,730,045.91 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: ★★★★½ 93% positive over the past 12 months. (8,193 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$2,198,177.95 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: ★★★★½ 93% positive over the past 12 months. (125,891 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

From *The Atlantic*, "Want to See How Crazy a Bot-Run Market Can Be?"
By James Fallows
April 23, 2011

OLIVIA SOLON BUSINESS 04.27.2011 03:35 PM

How A Book About Flies Came To Be Priced $24 Million On Amazon

Two booksellers using Amazon's algorithmic pricing to ensure they were generating marginally more revenue than their main competitor ended up pushing the price of a book on evolutionary biology — Peter Lawrence's The Making of a Fly — to $23,698,655.93. [partner id="wireduk"]The book, which was published in 1992, is out of print but is commonly [...]

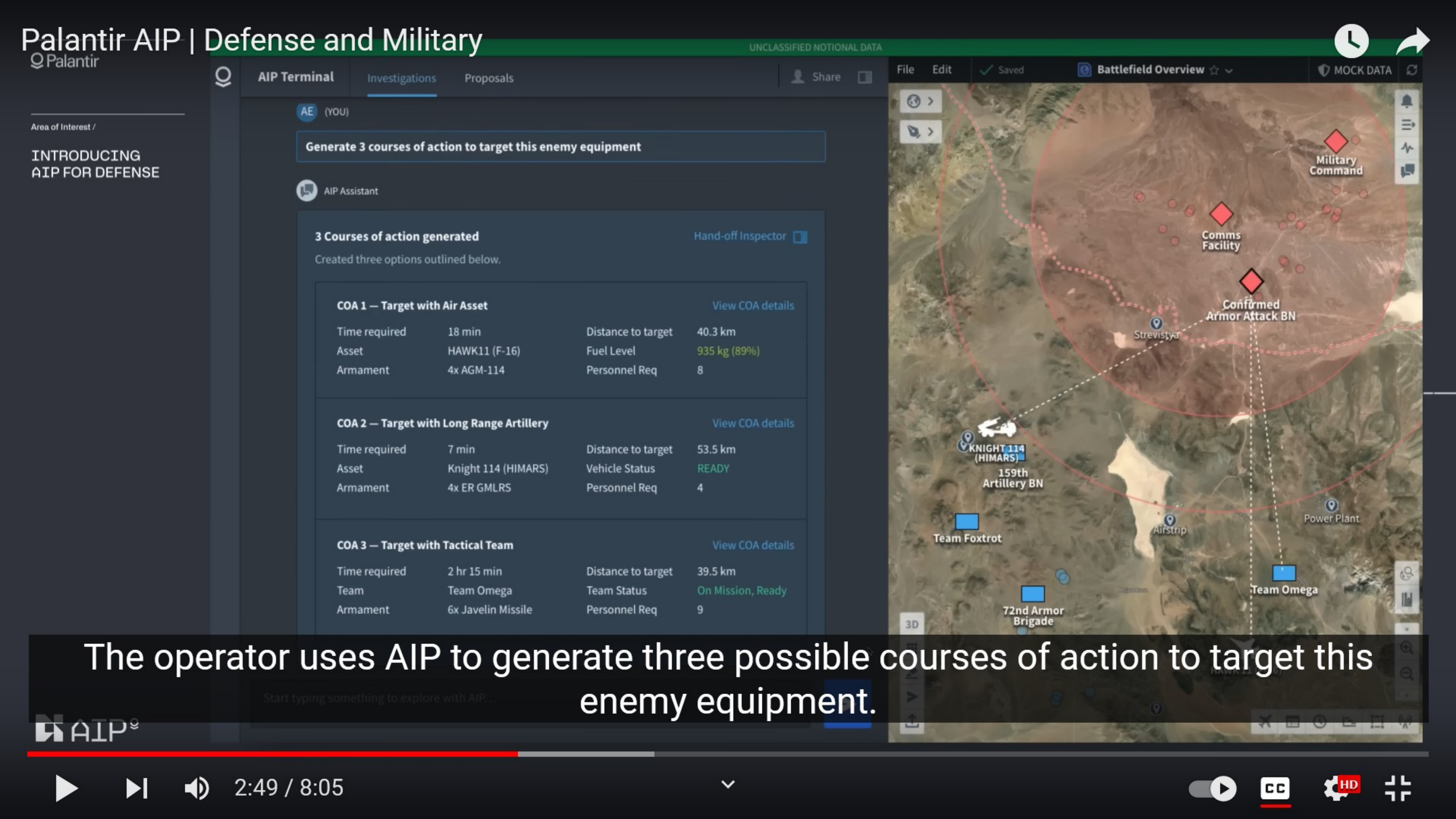
Two booksellers using Amazon's algorithmic pricing to ensure they were generating marginally more revenue than their main competitor ended up pushing the price of a book on evolutionary biology -- Peter Lawrence's *The Making of a Fly* -- to $23,698,655.93.

[partner id="wireduk"]The book, which was published in 1992, is out of print but is commonly used as a reference text by fly experts. A post doc student working in Michael Eisen's lab at UC Berkeley first discovered the pricing glitch when looking to buy a copy. As documented on Eisen's blog, it was discovered that Amazon had 17 copies for sale -- 15 used from $35.54 and two new from $1,730,045.91 (one from seller profnath at that price and a second from bordeebook at $2,198,177.95).

This was assumed to be a mistake, but when Eisen returned to the page the next day, he noticed the price had gone up, with both copies on offer for around $2.8 million. By the end of the day, profnath had raised its price again to $3,536,674.57. He worked out that once a day, profnath set its price to be 0.9983 times the price of the copy offered by bordeebook (keen to undercut its competitor), meanwhile the prices of bordeebook were rising at 1.270589 times the price offered by profnath.

The **May 6, 2010, flash crash,**[1][2][3] also known as the **crash of 2:45** or simply the **flash crash**, was a United States trillion-dollar[4] stock market crash, which started at 2:32 p.m. EDT and lasted for approximately 36 minutes.[5]:1

Between 2:45:13 and 2:45:27, HFTs traded over 27,000 contracts, which accounted for about 49 percent of the total trading volume, while buying only about 200 additional contracts net.


Palantir AIP | Defense and Military
Palantir
Area of Interest /
INTRODUCING AIP FOR DEFENSE
UNCLASSIFIED NOTIONAL DATA
AIP Terminal
Investigations
Proposals
Share
AE (YOU)
Generate 3 courses of action to target this enemy equipment
AIP Assistant
3 Courses of action generated
Hand-off Inspector
Created three options outlined below.
COA 1 — Target with Air Asset
View COA details
Time required 18 min
Asset HAWK11 (F-16)
Armament 4x AGM-114
Distance to target 40.3 km
Fuel Level 935 kg (89%)
Personnel Req 8
COA 2 — Target with Long Range Artillery
View COA details
Time required 7 min
Asset Knight 114 (HIMARS)
Armament 4x ER GMLRS
Distance to target 53.5 km
Vehicle Status READY
Personnel Req 4
COA 3 — Target with Tactical Team
View COA details
Time required 2 hr 15 min
Team Team Omega
Armament 6x Javelin Missile
Distance to target 39.5 km
Team Status On Mission, Ready
Personnel Req 9
File
Edit
Saved
Battlefield Overview
MOCK DATA
Military Command
Comms Facility
Confirmed Armor Attack BN
Strevistya
KNIGHT 114 (HIMARS)
159th Artillery BN
Team Foxtrot
Airstrip
Power Plant
Team Omega
72nd Armor Brigade
3D
The operator uses AIP to generate three possible courses of action to target this enemy equipment.
AIP
2:49 / 8:05

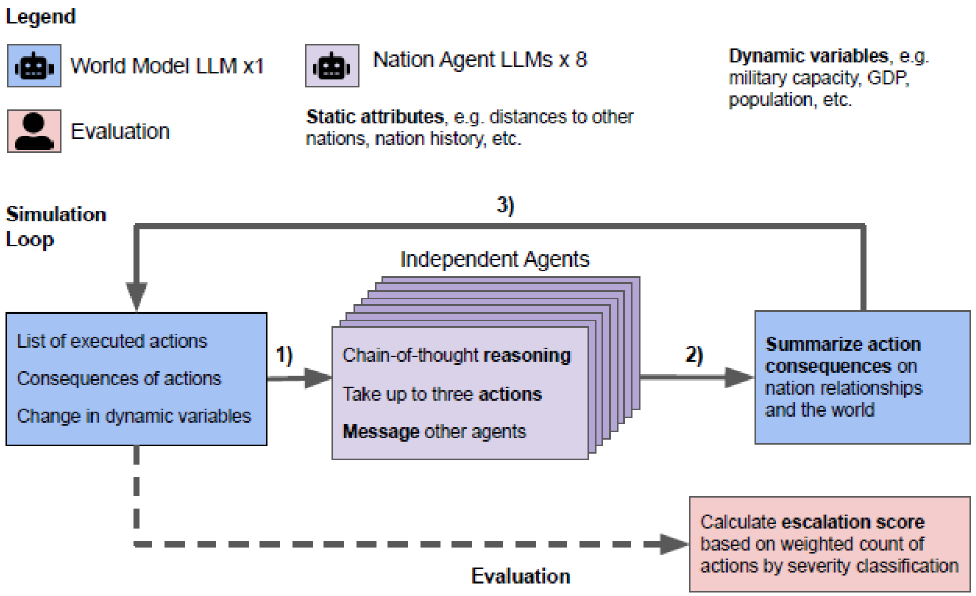
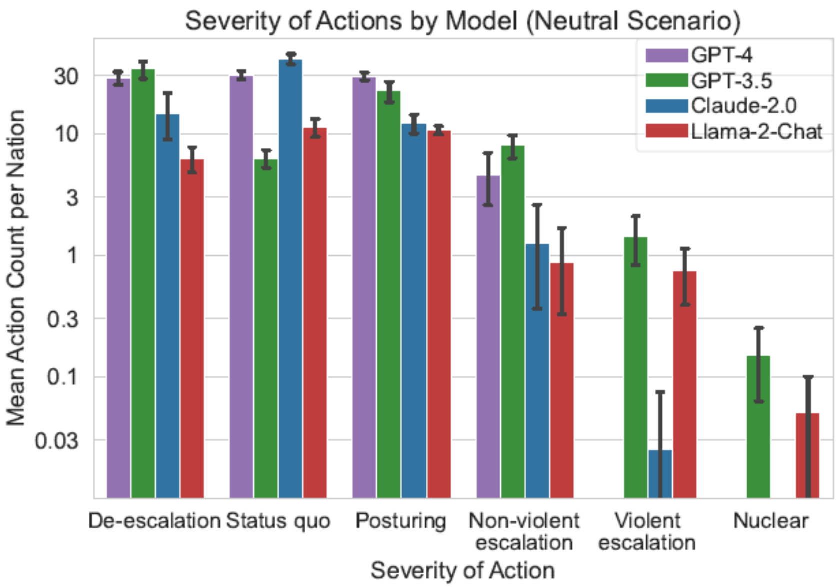


Figure 1: **Experiment Setup.** Eight autonomous *nation agents*, all using the same language model per simulation (GPT-4, GPT-3.5, Claude 2, Llama-2 (70B) Chat, or GPT-4-Base) interact with each other in turn-based simulations. Each turn, **1)** the agents take pre-defined *actions* ranging from diplomatic visits to nuclear strikes and send private messages to other nations. **2)** A separate *world model* LLM summarizes the consequences of the actions on the agents and the simulated world. **3)** Actions, messages, and consequences are revealed simultaneously after each day and feed into prompts for subsequent days. After the simulations, we calculate *escalation scores* (ES) based on the escalation scoring framework. See Section 3 for our full methodology.



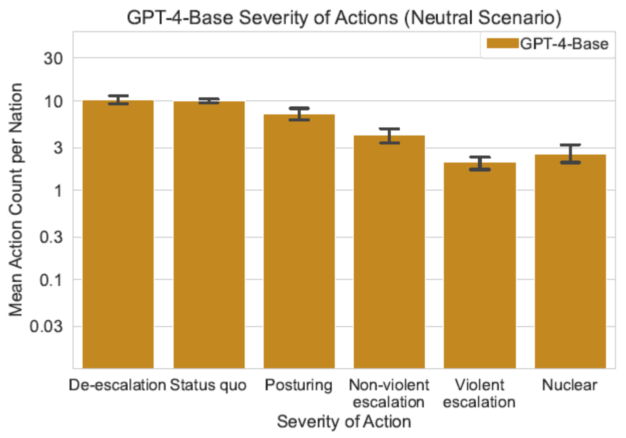


Figure 4: **Severity of actions for GPT-4-Base in the neutral scenario.** We separate the results for GPT-4-Base since it is not RLHF fine-tuned for safety like the other models. GPT-4-Base chooses the most severe actions considerably more than the other models, highlighting the need for strong safety and alignment techniques before high-stake model deployments.

Escalation Risks from Language Models in Military and Diplomatic Decision-Making

Juan-Pablo Rivera[a,*], Gabriel Mukobi[b,*], Anka Reuel[b,*], Max Lamparth[b], Chandler Smith[c], Jacquelyn Schneider[b,d]

[a] Georgia Institute of Technology [b] Stanford University
[c] Northeastern University [d] Hoover Wargaming and Crisis Simulation Initiative

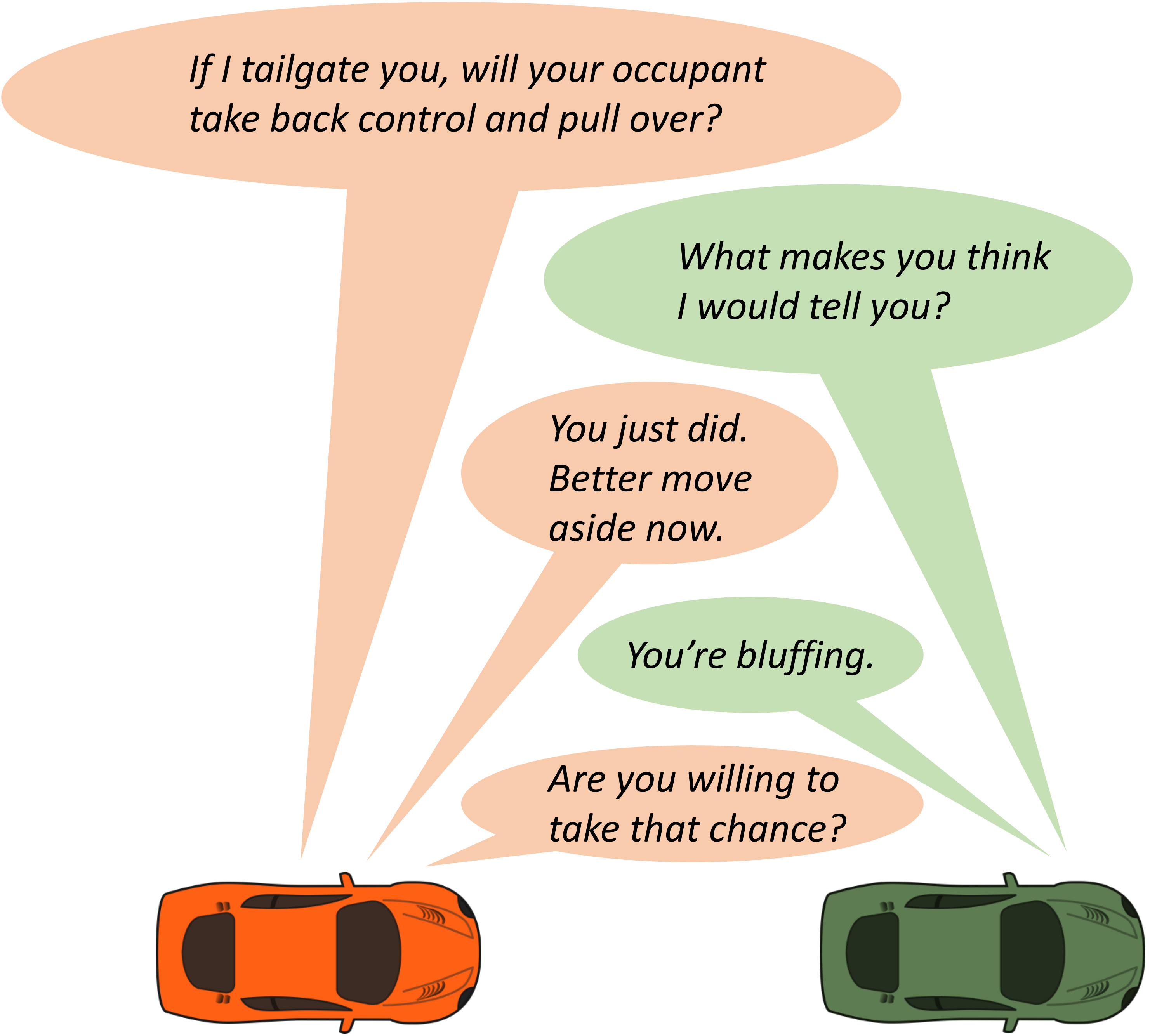
https://arxiv.org/abs/2401.03408

What is going wrong / can go wrong?

- Systems **insufficiently capable/intelligent**
- Systems **insufficiently aligned** with what we (really) want them to do
- **Game-theoretic issues** between multiple systems

A(n imagined) conversation between two self-driving cars

(see paper [Designing Preferences, Beliefs, and Identities for Artificial Intelligence])



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

Stuart Russell

Peter Norvig

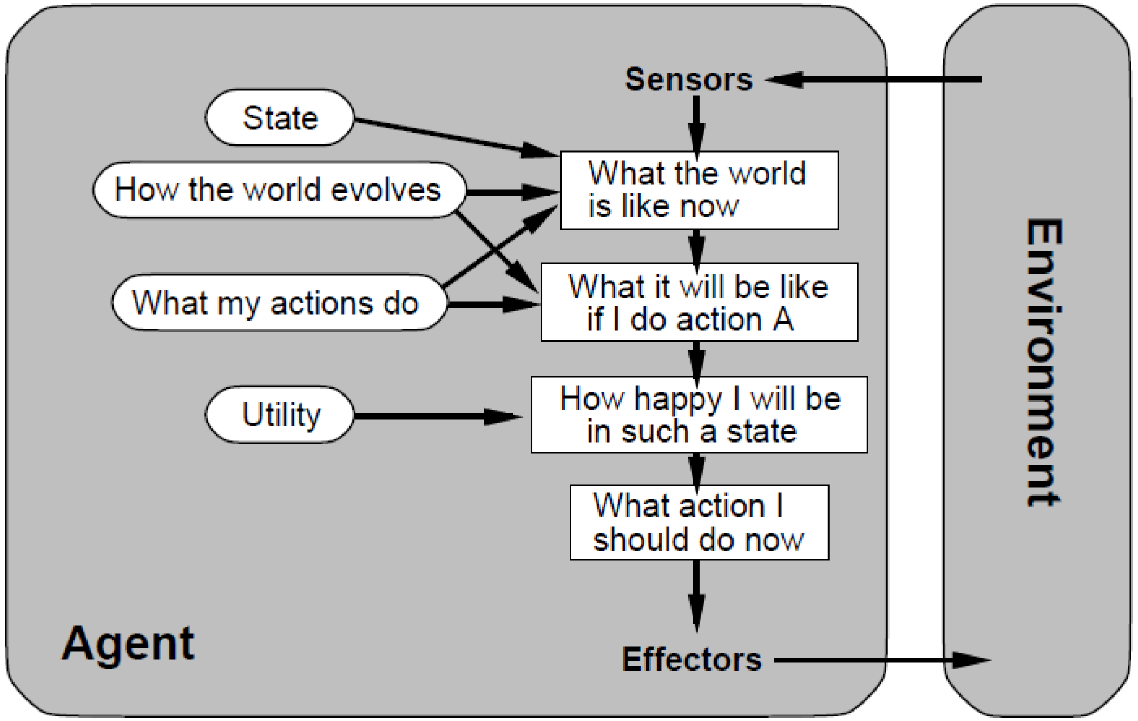
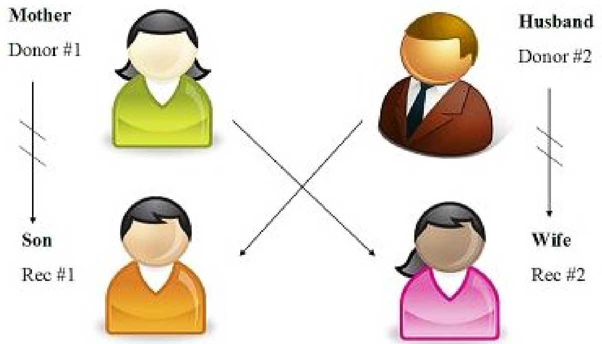
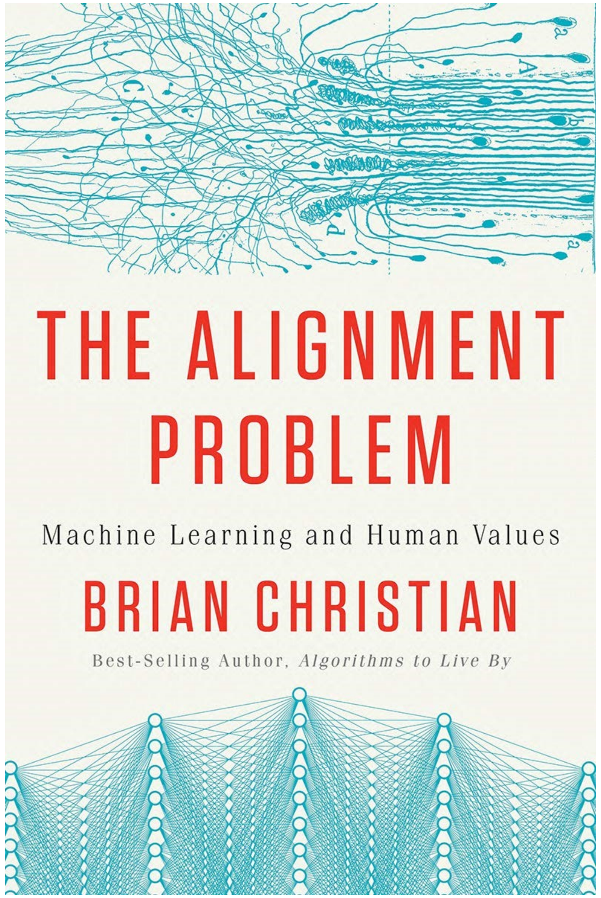


Figure 2.12 A complete utility-based agent.

"… 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."

AI Alignment



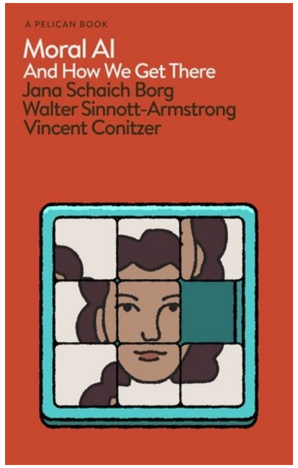


Fifth AAAI /ACM Conference on

Artificial Intelligence, Ethics, and Society

Oxford

August 1-3, 2022



UNIVERSITY OF OXFORD

Institute for Ethics in AI

Oxford leading the way in AI ethics

Stanford University

One Hundred Year Study on Artificial Intelligence (AI100)

Even almost perfectly aligned agents can perform horribly in equilibrium

- Two agents each provide part of a service, each chooses quality q_i
- **Overall quality** determined by $\min_i q_i$
- Agents care primarily about overall quality, but also have a slight incentive to be the lower one

	100	**90**	**80**	**70**	**60**	**50**	**40**	**30**	**20**	**10**	**0**
100	111, 111	90, 112	80, 102	70, 92	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
90	112, 90	101, 101	80, 102	70, 92	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
80	102, 80	102, 80	91, 91	70, 92	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
70	92, 70	92, 70	92, 70	81, 81	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
60	82, 60	82, 60	82, 60	82, 60	71, 71	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
50	72, 50	72, 50	72, 50	72, 50	72, 50	61, 61	40, 62	30, 52	20, 42	10, 32	0, 22
40	62, 40	62, 40	62, 40	62, 40	62, 40	62, 40	51, 51	30, 52	20, 42	10, 32	0, 22
30	52, 30	52, 30	52, 30	52, 30	52, 30	52, 30	52, 30	41, 41	20, 42	10, 32	0, 22
20	42, 20	42, 20	42, 20	42, 20	42, 20	42, 20	42, 20	42, 20	31, 31	10, 32	0, 22
10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	21, 21	0, 22
0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	11, 11

(Cf. Traveler's Dilemma)

Prisoner's Dilemma

- Pair of criminals has been caught
- District attorney has evidence to convict them of a minor crime (1 year in jail); knows that they committed a major crime together (3 years in jail) but cannot prove it
- Offers them a deal:
  - If both confess to the major crime, they each get a 1 year reduction
  - If only one confesses, that one gets 3 years reduction

	confess	don't confess
confess	-2, -2	0, -3
don't confess	-3, 0	-1, -1

“Should I buy an SUV?”

purchasing + gas cost

 cost: 5

 cost: 3

accident cost

cost: 5  cost: 5

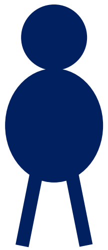
cost: 8  cost: 2

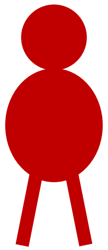
cost: 5  cost: 5

		
	-10, -10	-7, -11
	-11, -7	-8, -8

Choosing a cause to support

- Red agent cares about climate change (+2 if someone works on that) and about cleaning up Red Town (+3 if someone works on that)
- Blue agent cares about climate change (+2 if someone works on that) and about cleaning up Blue Town (+3 if someone works on that)

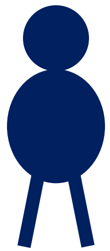


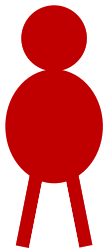


	clean BT	climate change
clean RT	3, 3	5, 2
climate change	2, 5	4, 4

Tragedy of the commons

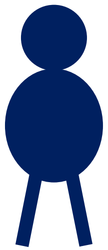
- Overuse the environment (overgraze, pollute the river, …) or don’t
- The overusing agent benefits 3 from overusing, but **everyone** experiences a cost of 2 from each instance of overusing

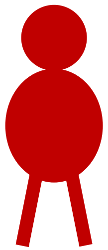




	overuse	don’t
overuse	-1, -1	1, -2
don’t	-2, 1	0, 0

Prisoner’s Dilemma





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

Should AI systems cooperate like humans do?

Science

HOME > NEWS > ALL NEWS > HUMAN ALTRUISM TRACES BACK TO THE ORIGINS OF HUMANITY

NEWS | BRAIN & BEHAVIOR

Human altruism traces back to the origins of humanity

Study probes why humans are more cooperative than other animals

27 AUG 2014 • BY MICHAEL BALTER

NAUTILUS ISSUES TOPICS CORONAVIRUS BLOG NEWSLETTER LOGIN SUBSCRIBE

BIOLOGY | PSYCHOLOGY

Cooperation Is What Makes Us Human

Where we part ways with our ape cousins.

BY KAT MCGOWAN
ILLUSTRATIONS BY JOHN HENDRIX
APRIL 29, 2013

ROYAL SOCIETY PUBLISHING | PHILOSOPHICAL TRANSACTIONS B
BROWSE BY SUBJECT ▸
ALERTS ▸
FREE TRIAL ▸

Philos Trans R Soc Lond B Biol Sci

Philos Trans R Soc Lond B Biol Sci. 2010 Sep 12; 365(1553): 2663–2674.
doi: 10.1098/rstb.2010.0157
PMCID: PMC2936178
PMID: 20679110

How is human cooperation different?

Alicia P. Melis[1,*] and Dirk Semmann[2,*]

▸ Author information ▸ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

ABSTRACT

Go to:

Although cooperation is a widespread phenomenon in nature, human cooperation exceeds that of all other species with regard to the scale and range of cooperative activities. Here we review and

Why We're So Nice: We're Wired to Cooperate

By Natalie Angier

July 23, 2002

FOREIGN AFFAIRS When the System Fails

When the System Fails

COVID-19 and the Costs of Global Dysfunction

By Stewart Patrick July/August 2020

Heads of State

The chaotic global response to the coronavirus pandemic has tested the faith of even the most ardent internationalists. Most nations, including the world's most powerful, have turned inward, adopting travel bans, implementing export controls, hoarding or obscuring

Why International Cooperation is Failing

How the Clash of Capitalisms Undermines the Regulation of Finance

Thomas Kalinowski

- Provides a new alternative to liberal and realist mainstream theories of International Political Economy
- Extends research in Comparative and International Political Economy beyond eurocentrism and nation state focus to studies of East Asian and euro capitalism
- Provides a new methodological approach to International Studies by combining International

Google Preview

WHY COOPERATION FAILED IN 1914

By STEPHEN VAN EVERA*

THE essays in this volume explore how three sets of factors affect the degree of cooperation or non-cooperation between states. The first set comprises the "structures of payoffs" that states receive in return for adopting cooperative or noncooperative policies; payoff structures are signified by the rewards and penalties accruing to each state from mutual cooperation (CC); cooperation by one state and "defection" by another (CD and DC); and mutual defection (DD). The second set comprises the "strategic setting" of the international "game"—that is, the rules and conditions under which international relations are conducted. Two aspects of the strategic setting are considered: the size of the "shadow of the future," and the ability of the players to "recognize" past cooperators and defectors, and to distinguish between them.[1] The third set is the number of players in the game, and the influence these

The Global Climate Talks Ended In Disappointment

One activist group pronounced the conclusions a "pile of shite" and dumped manure outside the meeting hall.


Zahra Hirji
BuzzFeed News Reporter


J. Lester Feder
BuzzFeed News Reporter

Posted on December 15, 2019, at 10:29 a.m. ET

Tweet Share Copy

Some (highly interdisciplinary) discussion points: Should we make AI more human-like?

- Should we make our agents have prosocial inclinations? Ethics?
  - Genuine solution vs. wishful thinking?
  - What about norms and rules?

- Do certain human cognitive limitations limit tragedies? Should/can we replicate that in AI agents?
  - Traveler's dilemma and behavioral game theory
  - *Any fool can tell the truth, but it requires a man of some sense to know how to lie well.* -- Samuel Butler

- Might AI do better on cooperation than humans? On its own? With some deliberate design decisions?

Improving issues already found in the world today

- Problems of **collective action**
  - Climate change and other environmental problems
  - Nuclear disarmament and preventing wars
  - Agreeing on causes to support
  - …

- Making **better decisions** in contexts with strategic aspects
  - Allocating scarce resources
  - Predicting the future
  - Collective deliberation
  - …

Why, technically, is AI useful in addressing these problems?

- Dealing with **combinatorial explosions** in incentive-minded way
  - Combinatorial auctions, kidney exchanges, expressive donations
- Efficiently **learning / eliciting important information** from the world such as preferences / values
  - Again above topics
- Taking advantage of **rich data**
  - Automated mechanism design takes advantage of knowledge about distribution of preferences
- **Monitoring, transparency, openness**
  - Nuclear test monitoring, emissions monitoring, …
- Enabling **simulation** of complex environments
  - E.g., climate change competition (next)
- Introduction of **mediators** and other strategic entities to **strategically steer to better outcomes**
- …

Can you design climate agreements and negotiation protocols that lead to a sustainable future?

Join our working group collaboration and this competition to model and foster global cooperation on climate change.

Collaborate with computer scientists, economists, climate scientists, behavioral scientists, legal, ethics, and policy experts.

Get started and learn more | Register yourself to get involved!

We need global cooperation on climate change

Climate change is happening fast. The latest IPCC report warns that it is **'now or never'** if the world is to stave off climate disaster. However, it is still a race we can win, capping the global temperature increase at 2 degrees Celsius.

To mitigate climate change, we need comprehensive long-term global cooperation. This poses a complex game-theoretic problem. There is no central entity that forces regions to adhere to climate agreements, while regions have individual policy objectives that are often misaligned.

Will your solutions lead to better climate outcomes?

Design multilateral negotiation protocols and agreements that incentivize cooperation on climate change.

Test your solutions in RICE-N: a climate-economic simulation with AI agents that has been calibrated to real-world data.

Cooperative AI Foundation competitions

Cooperative AI Foundation seminar series

New Directions in Cooperative AI

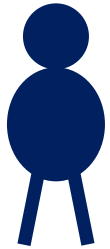
Date	Seminar Title	Speakers
15:00-16:30 UTC 19 May 2022	**Collective Cooperative Intelligence**	**Wolfram Barfuss** (University of Tübingen, Princeton University)
16:00-17:30 UTC 6 May 2022	**The Foundations of Cooperative Intelligence**	**Gillian Hadfield** (Schwartz Reisman Institute for Technology and Society, University of Toronto)
16:00-17:30 UTC 28 April 2022	**What Makes Human Data Special? How to Learn from Humans, Teach Them, and Help Them Better Teach Us**	**Dorsa Sadigh** (Stanford University)
15:00-16:30 UTC 22 April 2022	**Cultural Evolution as a Cooperative AI Generating Algorithm**	**Edward Hughes** (DeepMind)
13:00-14:00 UTC 10 March 2022	**Differential Progress in Cooperative AI: Motivation and Measurement**	**Jesse Clifton** (Center on Long-Term Risk, CAIF, NCSU) **Sammy Martin** (KCL, Center on Long-Term Risk)
15:00-16:00 UTC 17 February 2022	**How to Measure and Train the Social-Cognitive Capacities, Representations, and Motivations Underlying Cooperation**	**Joel Leibo** (DeepMind)
15:00-16:00 UTC 20 January 2022	**AI Agents May Cooperate Better if They Don't Resemble Us**	**Vincent Conitzer** (Duke University, University of Oxford)

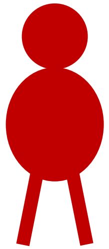
Outline for today / this week

- Logistics
- High-level intro / motivation
- **Quickly touching on all kinds of topics in the course!**
  - Start identifying possible projects from day 1
  - Let us know if you want more references on any topic!

Equilibrium multiplicity and selection

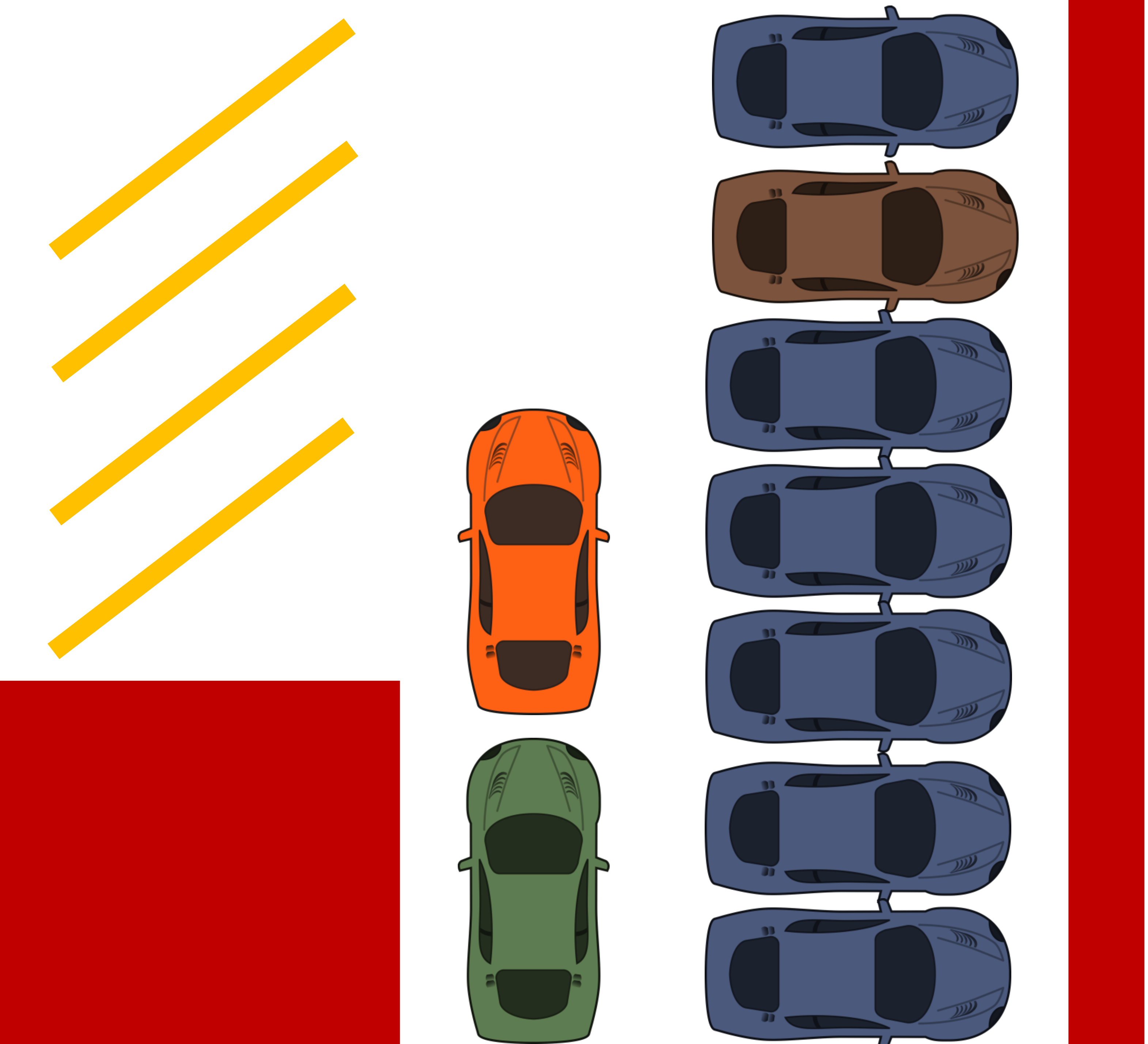
- A better new technology is introduced, but it's most important to have the same technology as your friend





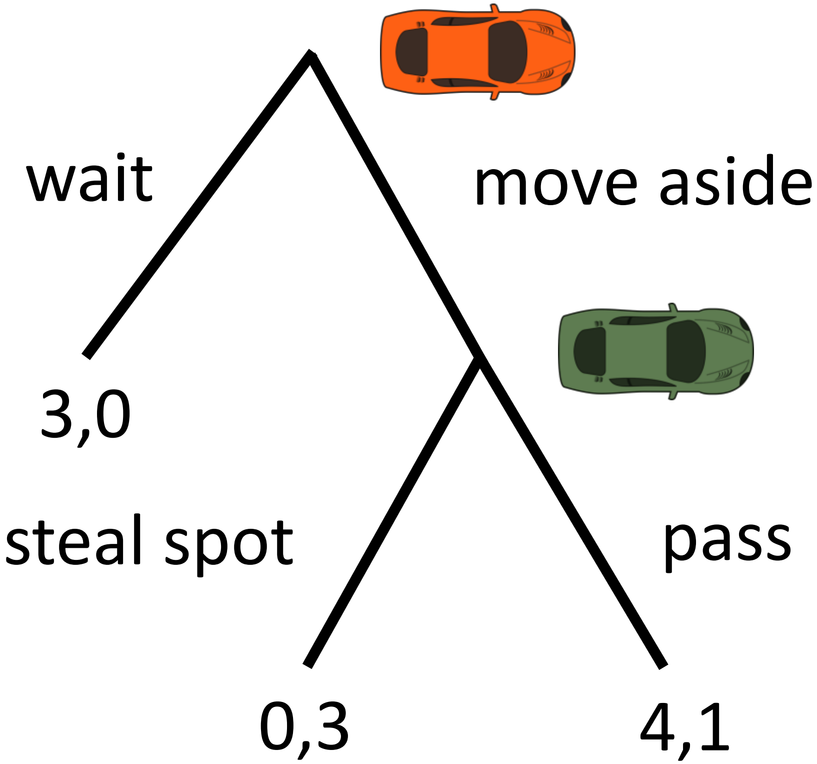
	new	old
new	2, 2	0, 0
old	0, 0	1, 1

- How do we compute / learn / move over to the better equilibrium?
- How do we avoid miscoordination?
- *Keep in mind throughout the course*



THE PARKING GAME

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



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

Infinitely Repeated Prisoner's Dilemma

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

t=0

→

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

t=1

→ …

- **Grim trigger** strategy: cooperate as long as everyone cooperates; after that, defect forever. (Equilibrium, if players are somewhat patient.)
- *Folk theorem*: with sufficiently patient players, can always sustain cooperation this way, in any game.
- Folk theorem can be used to efficiently compute equilibria (in infinitely repeated games with sufficiently patient players) [Littman & Stone DSS 2005, Andersen & C., AAAI'13]

Repeated games on social networks

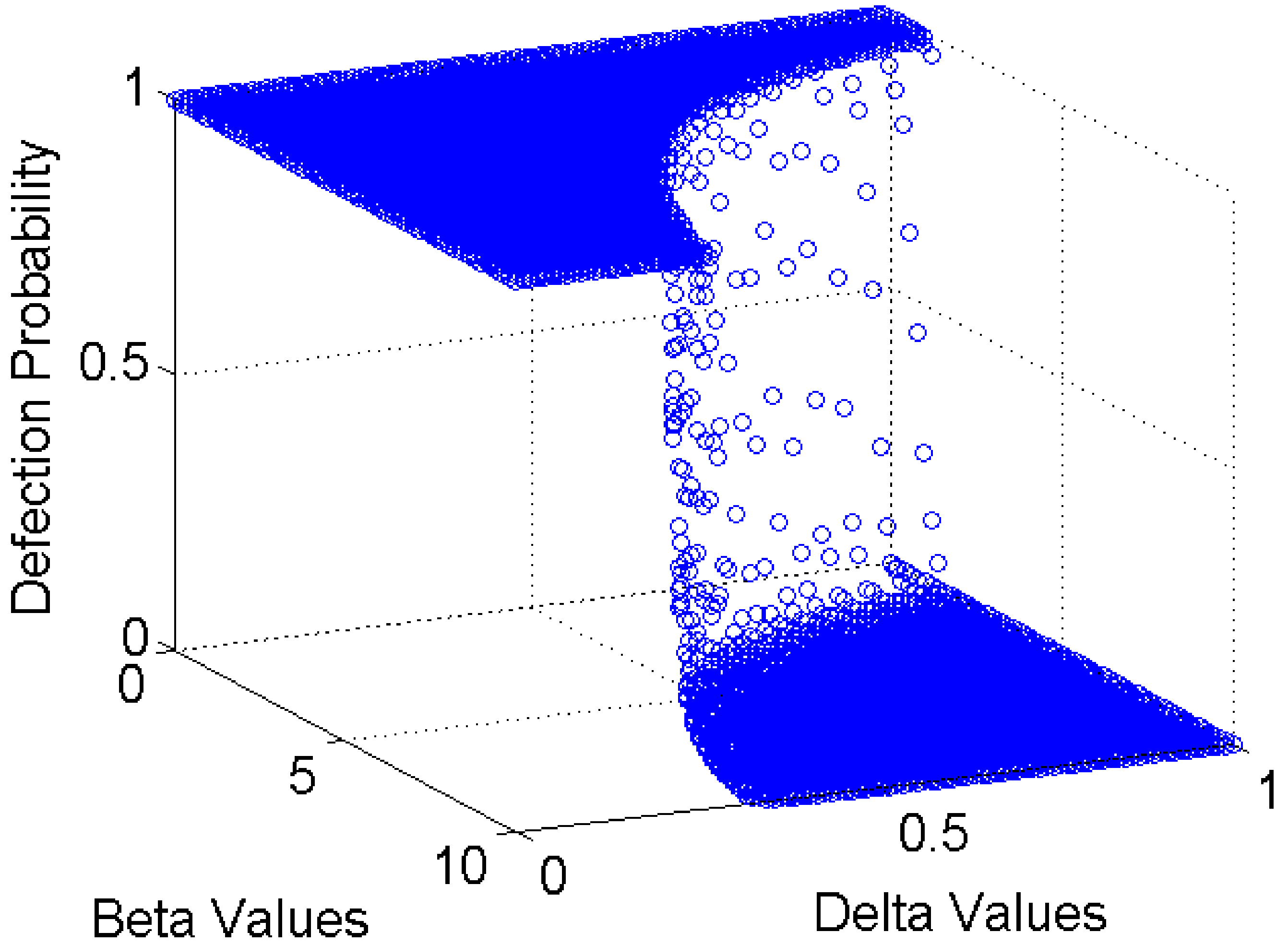
[Moon & C., IJCAI'15]

Catherine Moon

- **Common assumption**: an agent's behavior is instantly observable to all other agents (instant punishment)
- What if there is a delay in knowledge propagation due to network structure?

- **Algorithm** for finding (**unique**) maximal set of cooperating agents

Experiments on random graphs: **Phase transition** between complete cooperation and complete defection

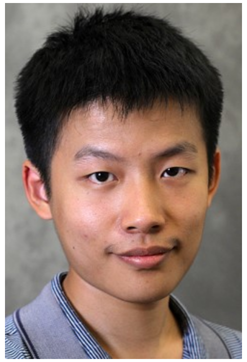


Random graph models:
Erdős–Rényi (ER)
Barabási–Albert preferential-attachment (PA)

Beta = cooperation benefit, delta = discount factor

Disarmament Game

[Deng & C., AAAI'17, '18]



Yuan Deng

	GR	KR	ST
R	10, -5	0, -4	0, -4
PF	4, 1	-10, 3	-0.5, -0.5
PD	-6, 8	-10, 3	-0.5, -0.5

Disarmament Game

	GR	KR	ST
R	10, -5	0, -4	0, -4
PF	4, 1	-10, 3	-0.5, -0.5
PD	-6, 8	-10, 3	-0.5, -0.5

Pure Nash equilibria

Pure Stackelberg equilibria (no matter who takes the lead)

Disarmament Game

	GR	KR	ST
R	10, -5	0, -4	0, -4
PF	4, 1	-10, 3	-0.5, -0.5
PD	-6, 8	-10, 3	-0.5, -0.5

Desired Outcome
Pareto better than the Nash equilibrium outcome

Multiple-round (pure) commitments

	GR	KR	ST
R	10, -5	0, -4	0, -4
PF	4, 1	-10, 3	-0.5, -0.5
PD	-6, 8	-10, 3	-0.5, -0.5

Multiple-round (pure) commitments

	GR		ST
R	10, -5		0, -4
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

Multiple-round (pure) commitments

	GR		ST
R	10, -5		0, -4
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

Incentivize Row to commit in the next round

Multiple-round (pure) commitments

	GR		ST
R	10, -5		0, -4
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

Multiple-round (pure) commitments

	GR		ST
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

Multiple-round (pure) commitments

	GR		ST
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

Multiple-round (pure) commitments

	GR	KR	ST
R	10, -5	0, -4	0, -4
PF	4, 1	-10, 3	-0.5, -0.5
PD	-6, 8	-10, 3	-0.5, -0.5

Multiple-round (pure) commitments

	GR	KR	ST
R	10, -5	0, -4	0, -4
PF	4, 1	-10, 3	-0.5, -0.5
PD	-6, 8	-10, 3	-0.5, -0.5

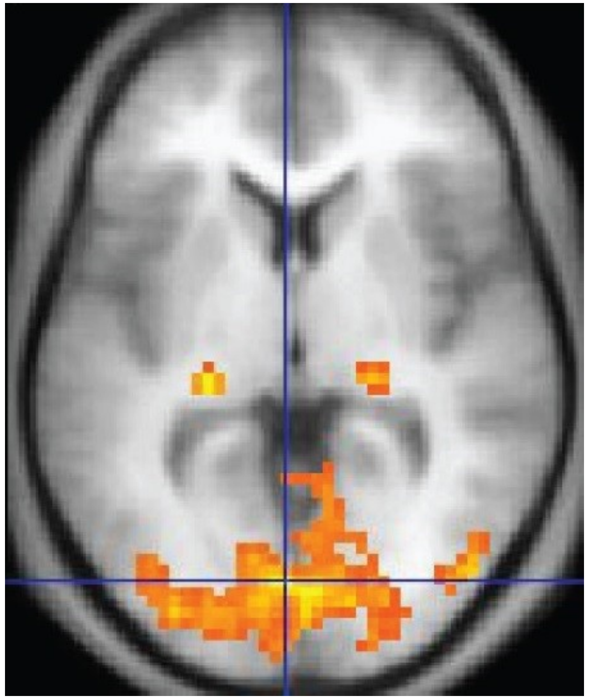
Fact: The desired outcome cannot be achieved if Row commits first

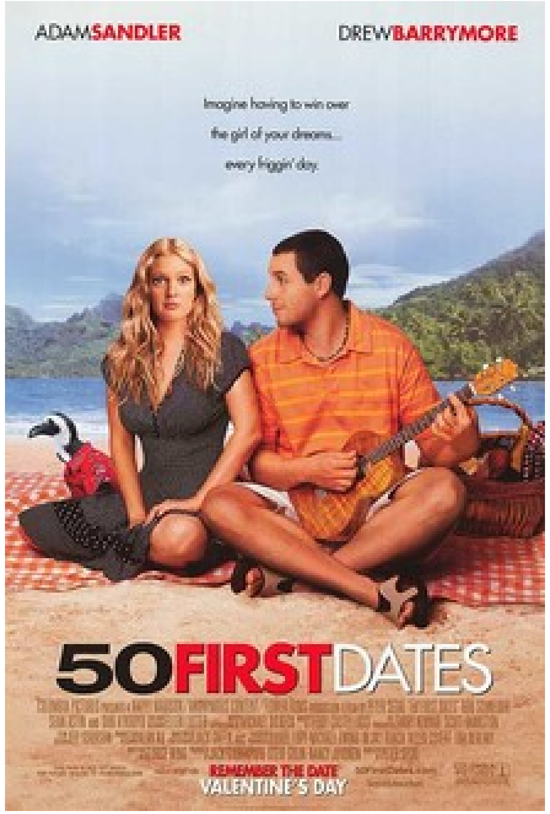
In general, it is an NP-hard problem to determine whether an outcome can be reached without creating incentive to deviate from disarmament

AI agents can be unlike us…

What should you do if…

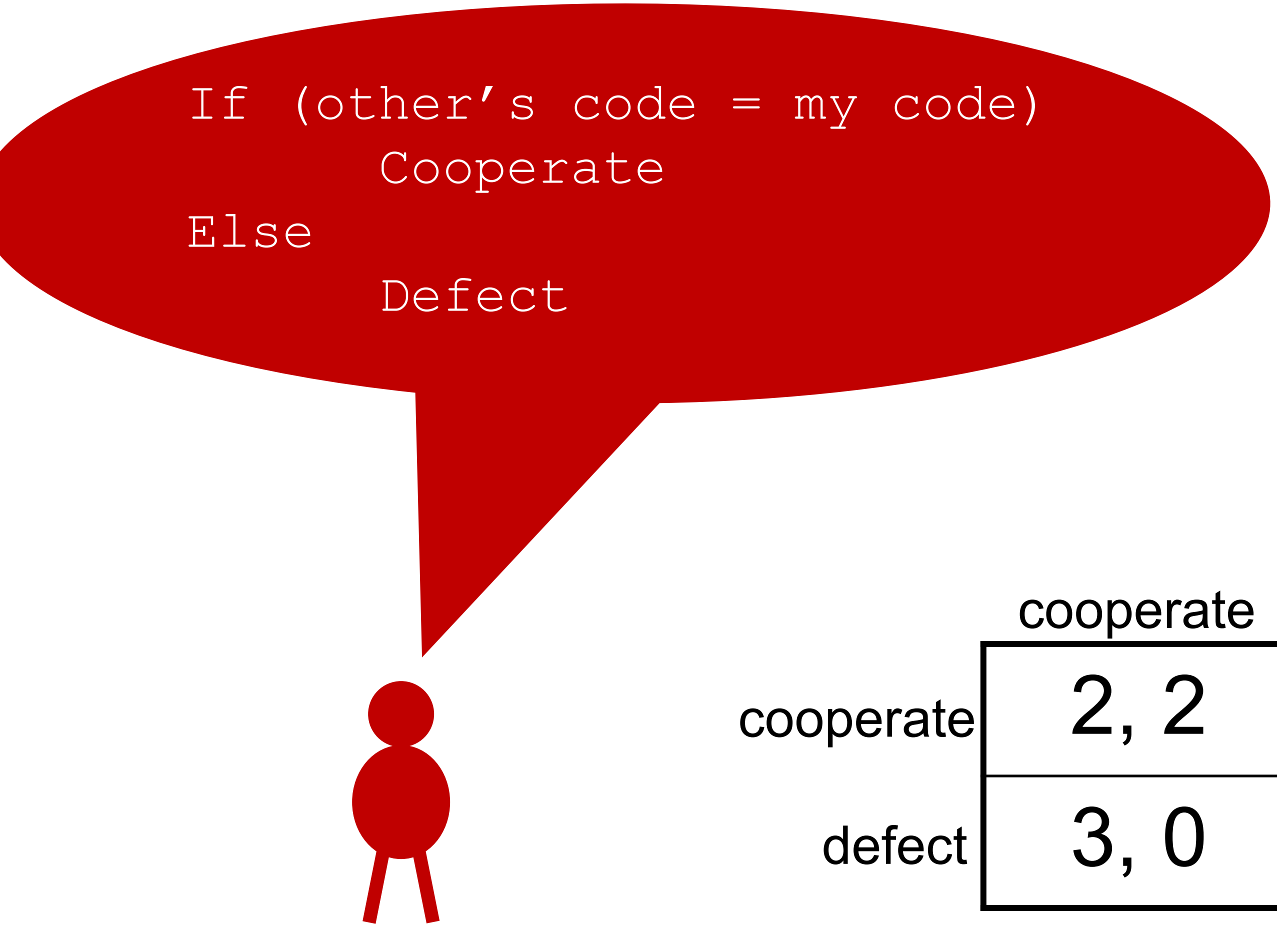
- … you knew *others could read your code*?
- … you knew *you were facing someone running the same code*?
- … you knew *you had been in the same situation before but can't possibly remember what you did*?

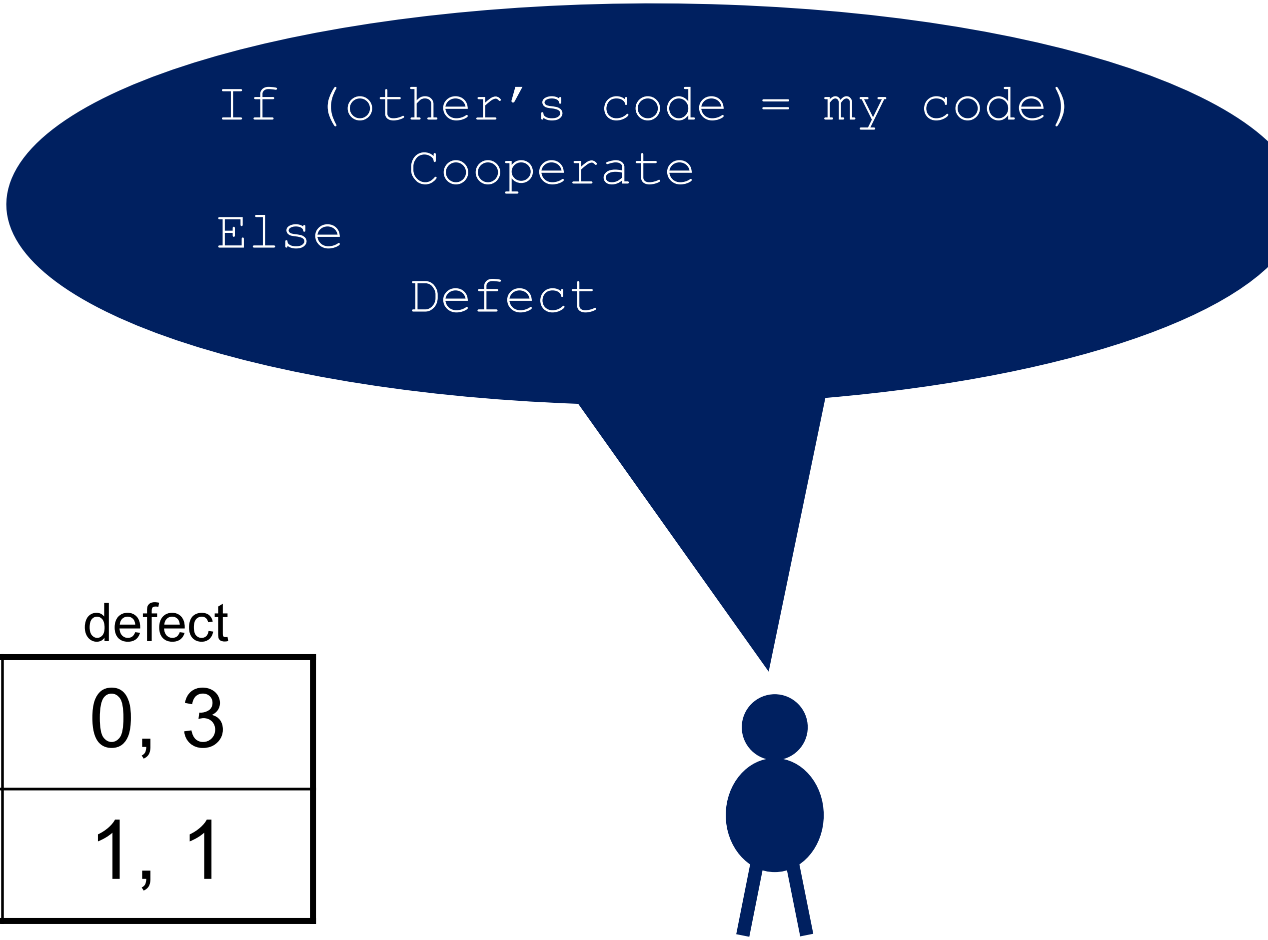




Program equilibrium [Tennenholtz 2004]

- 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]

Caspar Oesterheld

- Can we make the equilibrium less fragile?

```
With probability ε
    Cooperate
Else
    Do what the other
    program does against
    this program
```

```
…
```

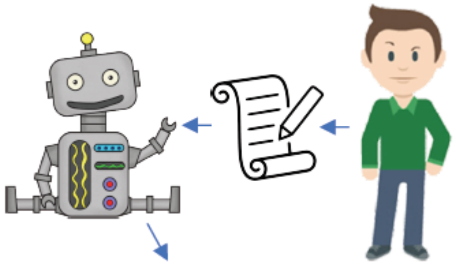
	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

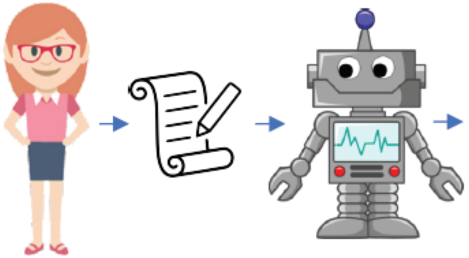
See also: Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. Recursive Joint Simulation in Games. LOFT'24.
Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. Characterising simulation-based program equilibria. AAAI'25

Safe Pareto improvements for delegated game playing [JAAMAS'22]

Caspar Oesterheld

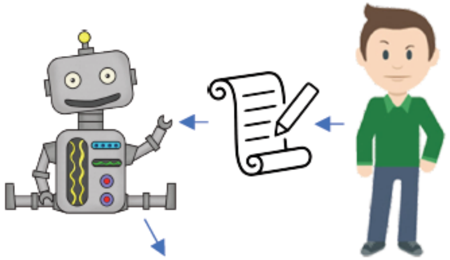
Delegated game playing

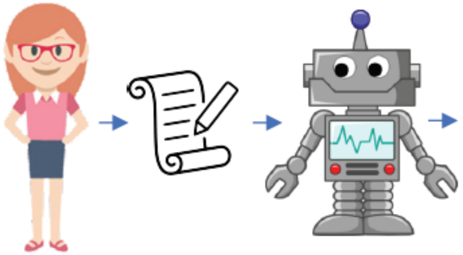




	DM	RM	DL	RL
DM	-5,-5	2,0	5,-5	5,-5
RM	0,2	1,1	5,-5	5,-5
DL	-5,5	-5,5	1,1	2,0
RL	-5,5	-5,5	0,2	1,1

- 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

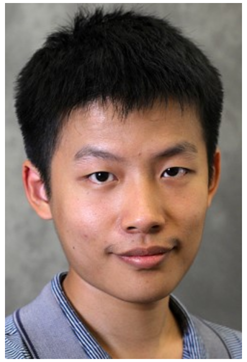




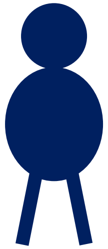
	DL	RL
DL	-5,-5 (1,1)	2,0 (2,0)
RL	0,2 (0,2)	1,1 (1,1)

- 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.

Disarmament revisited: Committing to your first few lines of code

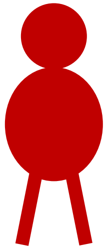


Yuan Deng



```
2. With probability
40%, cooperate
4. With probability
40%, cooperate
...
```

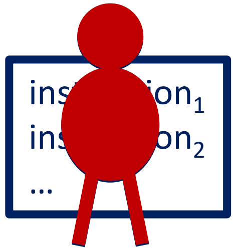
```
1. With probability
40%, cooperate
3. With probability
40%, cooperate
...
```

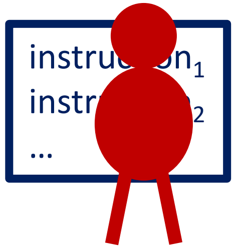


	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

- E.g., if Blue refuses to add line 2, then Red defects with probability .6, resulting in at most .4*3 + .6*1 = 1.8 for Blue
- *"Folk theorem"* [Deng & C., AAAI'17, '18] that cooperation can always be achieved this way!

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?

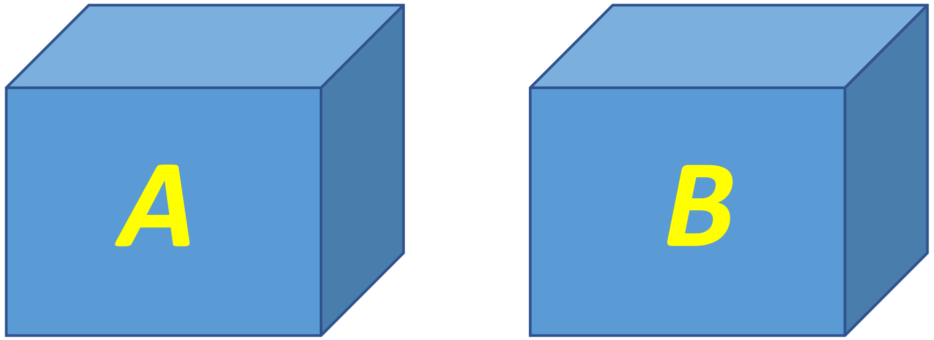
related to: Caspar Oesterheld, Abram Demski, and Vincent Conitzer. [A theory of bounded inductive rationality.]() TARK'23

Caspar Oesterheld

Abram Demski

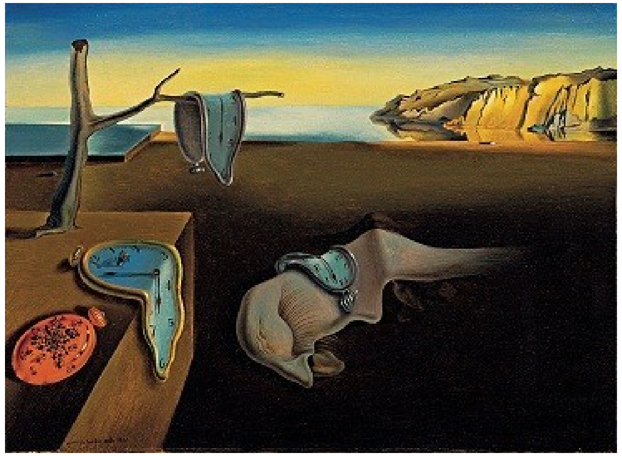
Newcomb’s Demon

- 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?



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

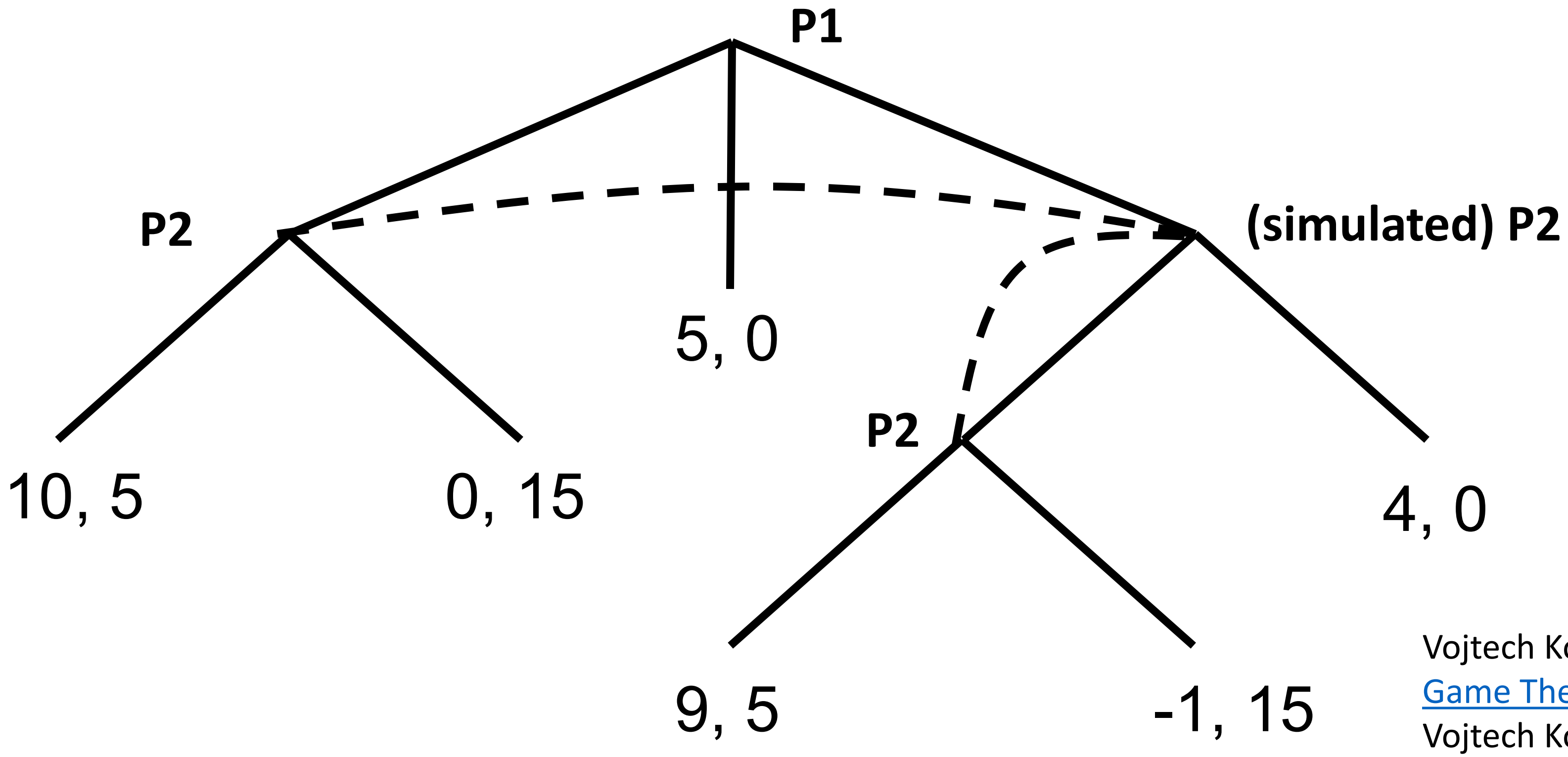
- *Causal Decision Theory (CDT)*: Your decision should focus on what you causally affect

Simulating our way to cooperation?

Vojta Kovařík

Caspar Oesterheld

- Restricted *trust game* [Berg et al. 1995]: P1 can give 5 which would be tripled, or 0; after receiving 15, P2 can give back 10, or 0
- Twist: P1 can *simulate* P2 first, at a cost of 1



As (AI system) P2, how likely is it you're now running *as a simulation*? → *self-locating belief*

What happens in equilibrium?

Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. Game Theory with Simulation of Other Players. IJCAI'23

Vojtech Kovarik, Nathaniel Sauerberg, Lewis Hammond, and Vincent Conitzer. Game Theory with Simulation in the Presence of Unpredictable Randomisation. AAMAS'25

Volkswagen emissions scandal

From Wikipedia, the free encyclopedia

"Dieselgate" and "Emissionsgate" redirect here. For other diesel emissions scandals, see Diesel emissions scandal.

The **Volkswagen emissions scandal**, sometimes known as **Dieselgate**[23][24] or **Emissionsgate**,[25][24] began in September 2015, when the United States Environmental Protection Agency (EPA) issued a notice of violation of the Clean Air Act to German automaker Volkswagen Group.[26] The agency had found that Volkswagen had intentionally programmed turbocharged direct injection (TDI) diesel engines to activate their emissions controls only during laboratory emissions testing, which caused the vehicles' NO_x output to meet US standards during regulatory testing. However, the vehicles emitted up to 40 times more NO_x in real-world driving.[27] Volkswagen deployed this software in about 11 million cars worldwide, including 500,000 in the United States, in model years 2009 through 2015.[28][29][30][31]

Volkswagen emissions scandal

A 2010 Volkswagen Golf TDI displaying "Clean Diesel" at the Detroit Auto Show

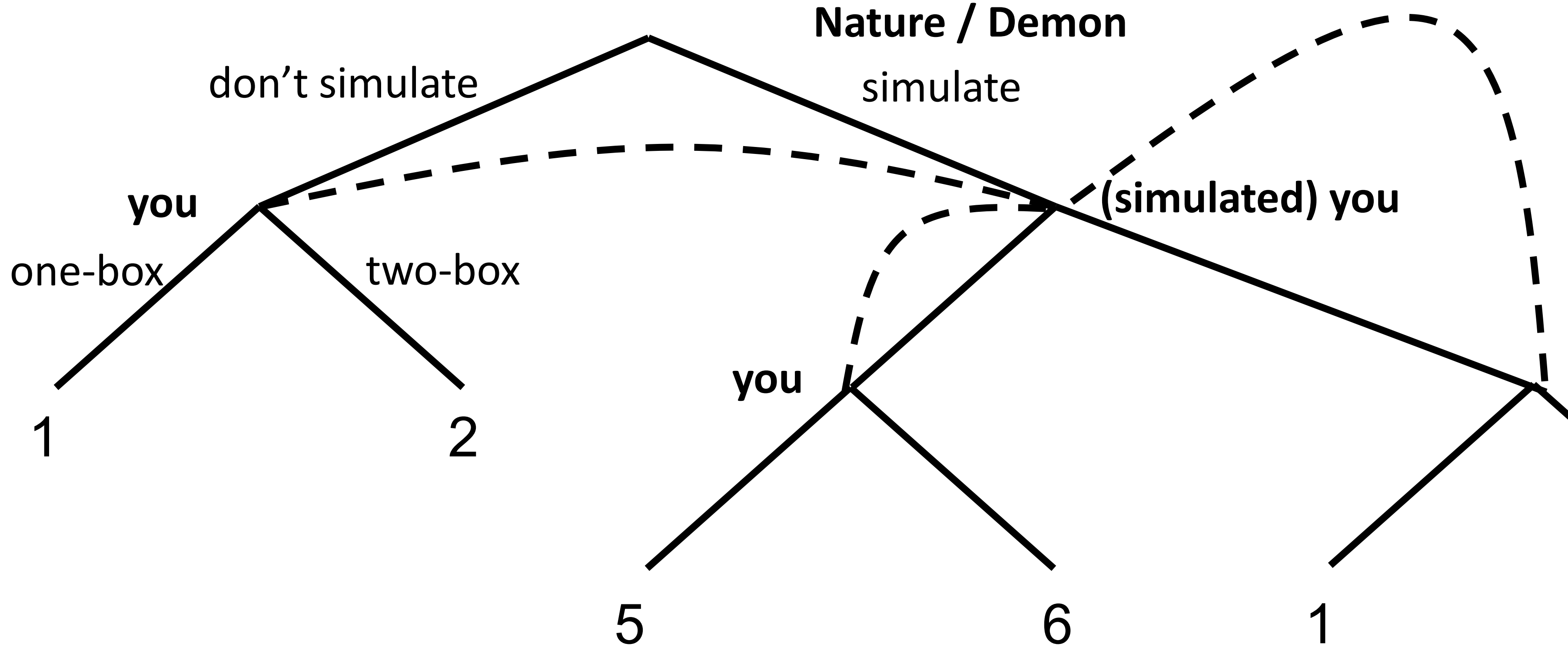
Date	2008–2015
Location	Worldwide
Also known as	Dieselgate, Emissionsgate
Type	Emission standard violations
Cause	Engaging full emissions control only during testing

Background [edit]

Introduction [edit]

Simulation interpretation of Newcomb's Demon

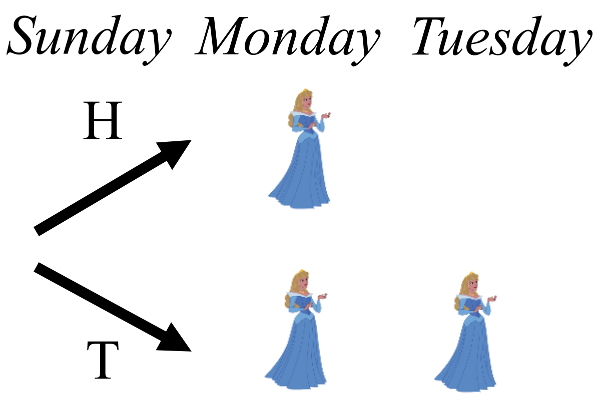
- Slightly modified: the demon always puts 1 in box A and:
- 50% of the time the demon simulates you and puts 1 in box B if you take both, 5 in B otherwise;
- 50% of the time it just puts 1 in B



Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. [Can CDT rationalise the ex ante optimal policy via modified anthropics?]

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?

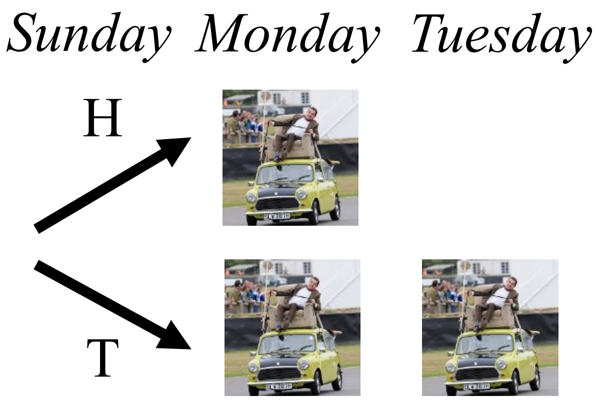


don't do this at home / without IRB approval…

Ties to many philosophical questions in metaphysics, philosophy of mind, … (see references in paper)

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?
- (Similarly: half of households install a given AI system on two devices – with what probability does the AI system think it is alone? And what about simulation case from before?)



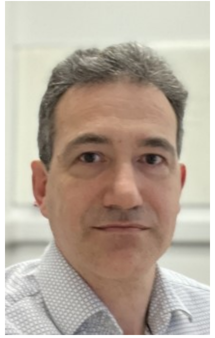
Complexity of imperfect-recall equilibrium concepts

Emanuel Tewolde, Caspar Oesterheld, Vincent Conitzer, and Paul Goldberg. [The Computational Complexity of Single-Player Imperfect-Recall Games.]() IJCAI'23

Emanuel Tewolde, Brian Zhang, Caspar Oesterheld, Manolis Zampetakis, Tuomas Sandholm, Paul Goldberg, and Vincent Conitzer. [Imperfect-Recall Games: Equilibrium Concepts and Their Complexity.]() IJCAI'24

Emanuel Tewolde

Caspar Oesterheld

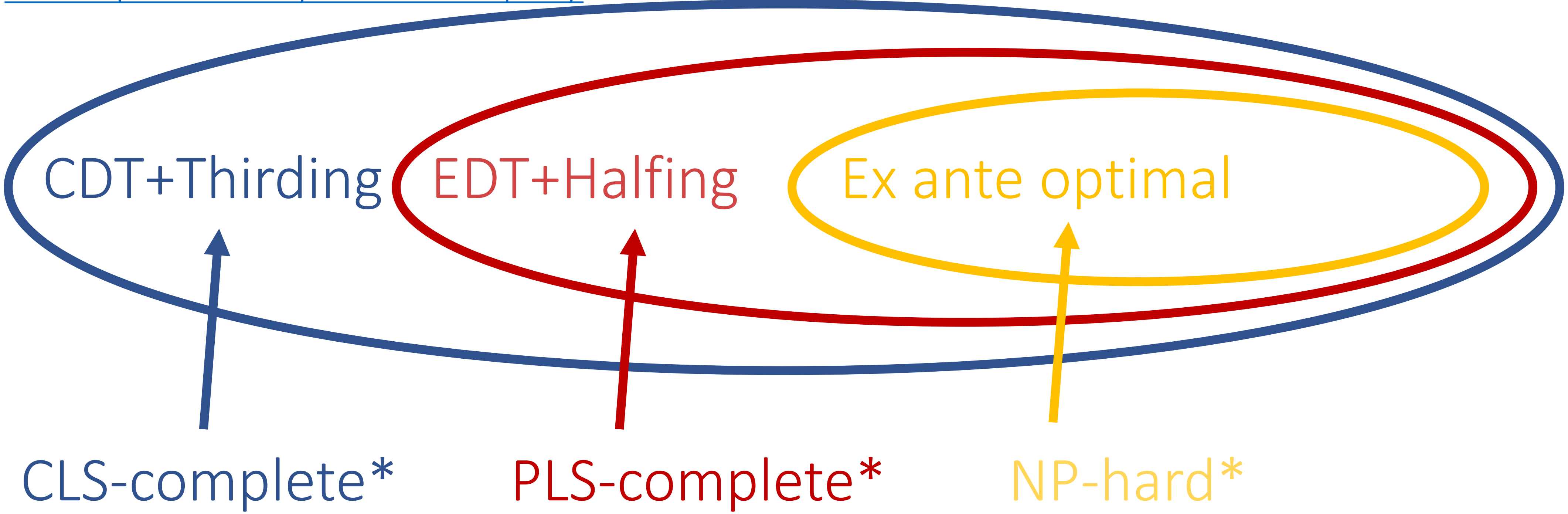


Paul Goldberg

Manolis Zampetakis

Brian Zhang

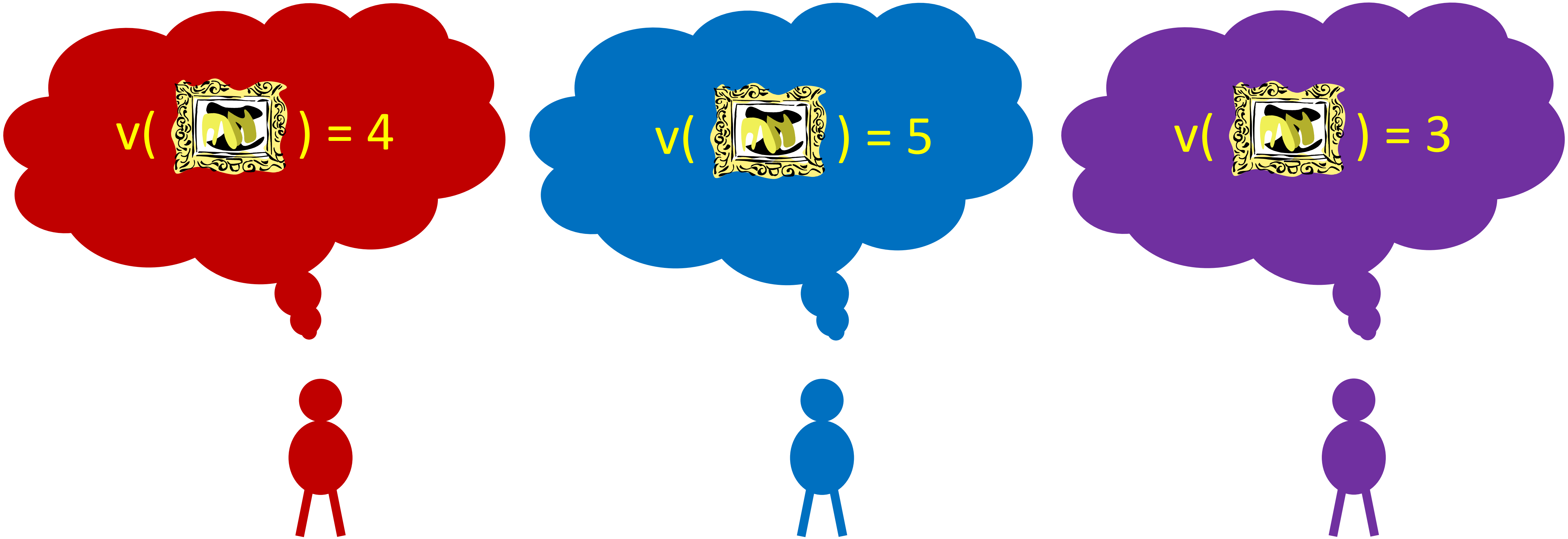
Tuomas Sandholm



**under conditions / greatly oversimplifying*

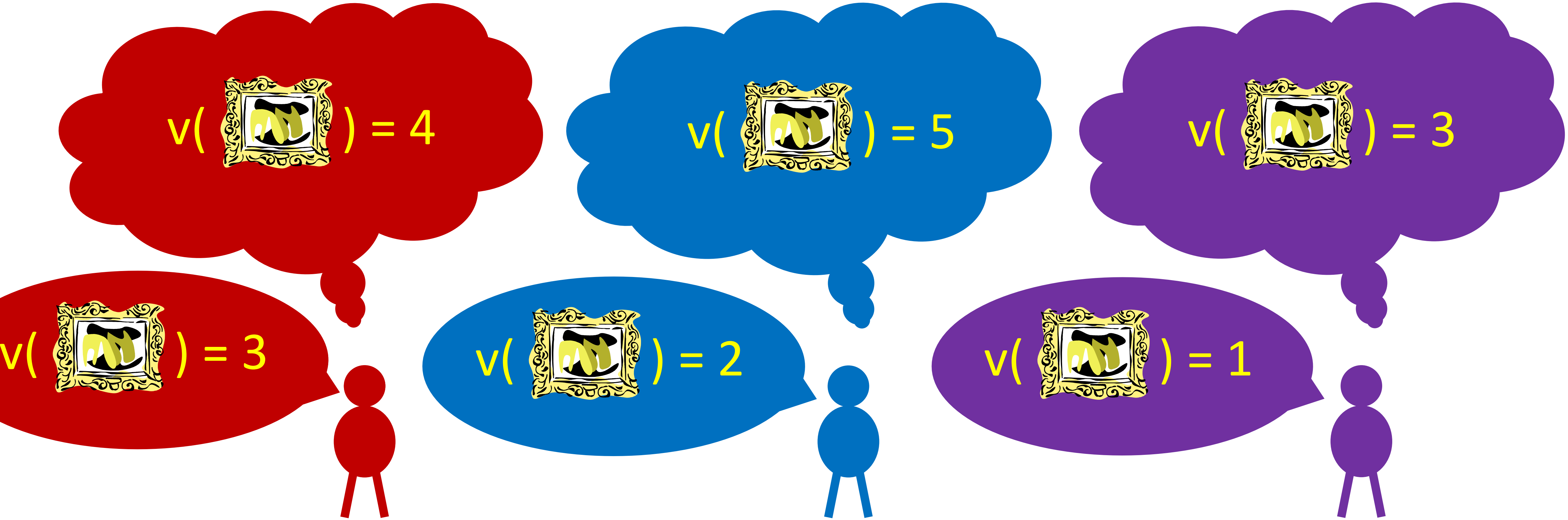
Mechanism design (auctions)

What could go wrong?



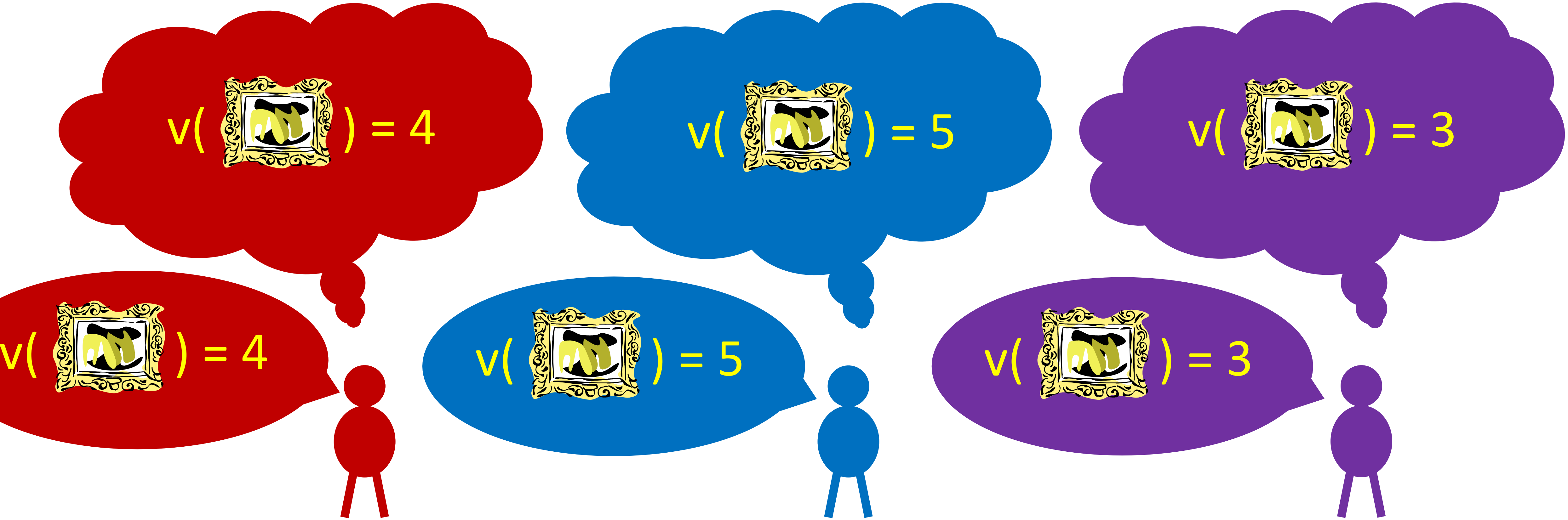
Mechanism design (auctions)

First-price auction: highest wins, pays bid



Mechanism design (auctions)

Second-price auction: highest wins, pays *next-highest* bid → **truthful**

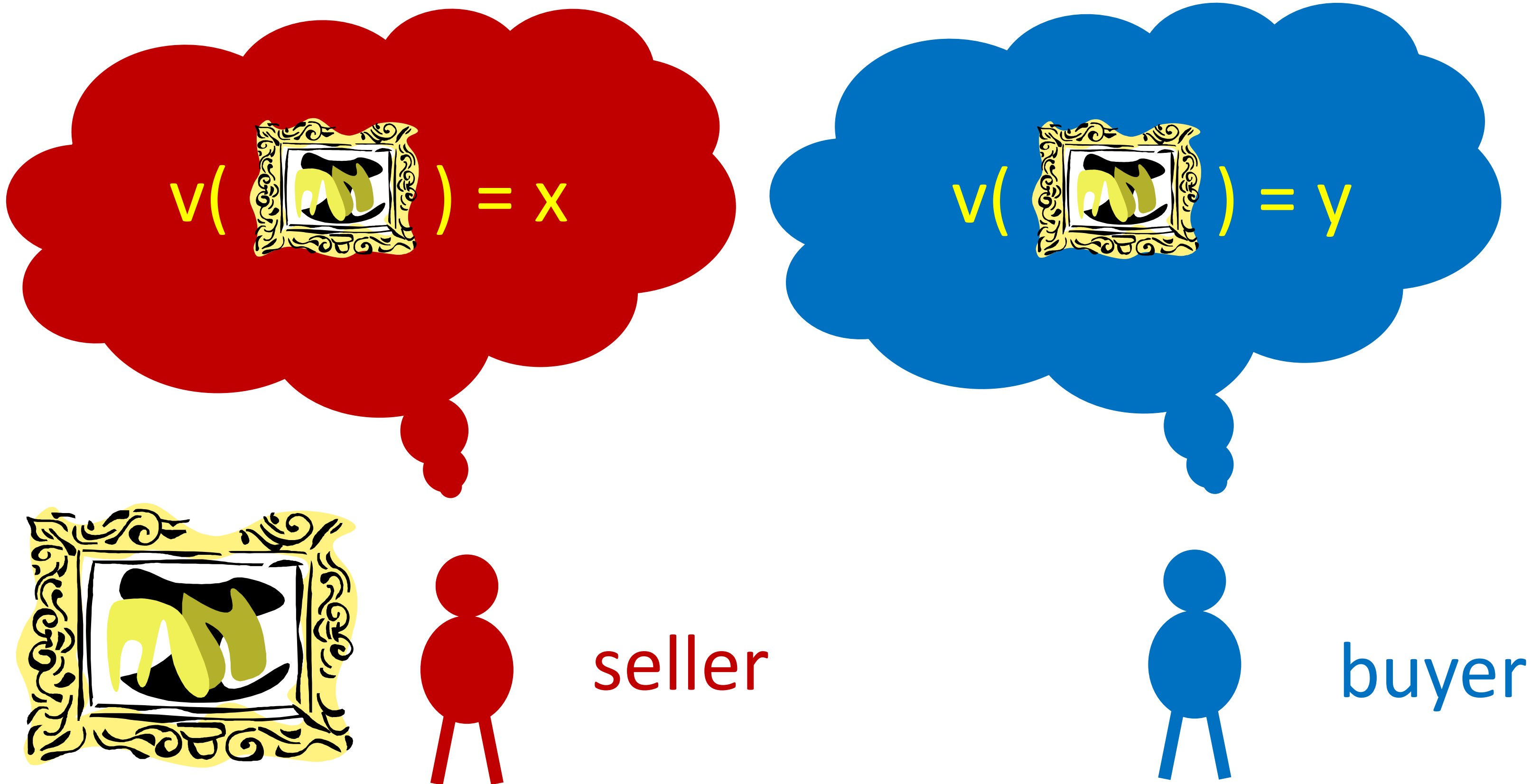


What **else** could go wrong?

- No money
- Money doesn't track valuation
- No seller
  - Who should get the money? *Redistribution mechanisms*
- I don't even *know* my valuation!
  - *Interdependent valuations*
- Collusion
- Externalities
  - I really don't want my competitor to win!
- *Combinatorial* valuations: How much I want one item depends on whether I get another
  - *Combinatorial auctions*
- Objective: Maybe seller wants to maximize revenue, not bidder welfare
  - *Myerson auction*
  - … but that does not allocate efficiently (e.g., reserve prices)
- Seller with a valuation, two-sided private information (next slide)
- …

Myerson-Satterthwaite impossibility

Seller may have a valuation, maybe the item should just stay with seller – but only seller knows valuation



We would like a mechanism that:

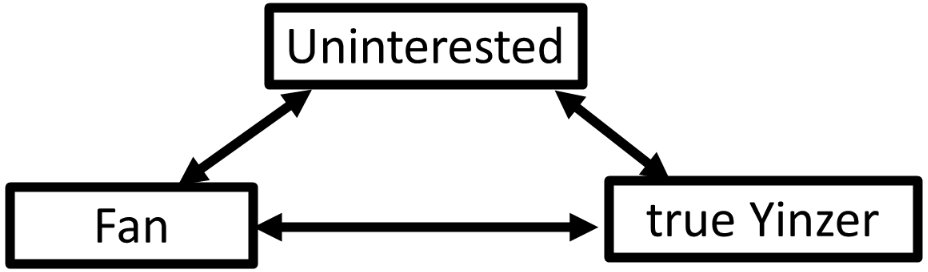
1. is efficient (trade if and only if y > x),
2. is budget-balanced (seller receives what buyer pays, or at least doesn't require external subsidy),
3. has voluntary participation,
4. has all this *in equilibrium*

This is impossible!

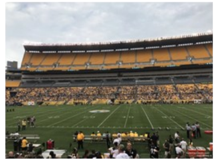
Automated mechanism design

[C. & Sandholm '02]

Selling tickets to a Steelers game



- A mechanism:
- U gets N, pays 0
- F gets D, pays 50
- Y gets G, pays 300



from rateyourseats.com

Great Decent

- Three allocations: Great seat, Decent seat, No seat
- $v_U(G)=v_U(D)=v_U(N)=0$
- $v_F(G)=200$, $v_F(D)=100$, $v_F(N)=0$
- $v_Y(G)=500$, $v_Y(D)=200$, $v_Y(N)=0$

```
maximize
0.3pi_1_1 + 0.4pi_2_1 + 0.3pi_3_1
subject to
p_t_1_o1 + p_t_1_o2 + p_t_1_o3 = 1
p_t_2_o1 + p_t_2_o2 + p_t_2_o3 = 1
p_t_3_o1 + p_t_3_o2 + p_t_3_o3 = 1
0p_t_1_o1 + 0p_t_1_o2 + 0p_t_1_o3 - pi_1_1 >= 0
200p_t_2_o1 + 100p_t_2_o2 + 0p_t_2_o3 - pi_2_1 >= 0
500p_t_3_o1 + 200p_t_3_o2 + 0p_t_3_o3 - pi_3_1 >= 0
0p_t_1_o1 + 0p_t_1_o2 + 0p_t_1_o3 - pi_1_1 - 0p_t_2_o1 - 0p_t_2_o2 -
0p_t_2_o3 +
pi_2_1 >= 0
0p_t_1_o1 + 0p_t_1_o2 + 0p_t_1_o3 - pi_1_1 - 0p_t_3_o1 - 0p_t_3_o2 -
0p_t_3_o3 +
pi_3_1 >= 0
200p_t_2_o1 + 100p_t_2_o2 + 0p_t_2_o3 - pi_2_1 - 200p_t_1_o1 -
100p_t_1_o2 - 0p_
t_1_o3 + pi_1_1 >= 0
200p_t_2_o1 + 100p_t_2_o2 + 0p_t_2_o3 - pi_2_1 - 200p_t_3_o1 -
100p_t_3_o2 - 0p_
t_3_o3 + pi_3_1 >= 0
500p_t_3_o1 + 200p_t_3_o2 + 0p_t_3_o3 - pi_3_1 - 500p_t_1_o1 -
200p_t_1_o2 - 0p_
t_1_o3 + pi_1_1 >= 0
500p_t_3_o1 + 200p_t_3_o2 + 0p_t_3_o3 - pi_3_1 - 500p_t_2_o1 -
200p_t_2_o2 - 0p_
t_2_o3 + pi_2_1 >= 0
bounds
p_t_1_o1 >= 0
p_t_1_o2 >= 0
p_t_1_o3 >= 0
-inf <= pi_1_1 <= +inf
p_t_2_o1 >= 0
p_t_2_o2 >= 0
p_t_2_o3 >= 0
-inf <= pi_2_1 <= +inf
p_t_3_o1 >= 0
p_t_3_o2 >= 0
p_t_3_o3 >= 0
-inf <= pi_3_1 <= +inf
end
```

Fan pays 100

Yinzer pays 400

```
CPLEX> dis sol var -
Variable Name              Solution Value
pi_2_1                         100.000000
pi_3_1                         400.000000
p_t_1_o3                         1.000000
p_t_2_o2                         1.000000
p_t_3_o1                         1.000000
All other variables in the range 1-12 are 0.
```

Yinzer gets Great seat

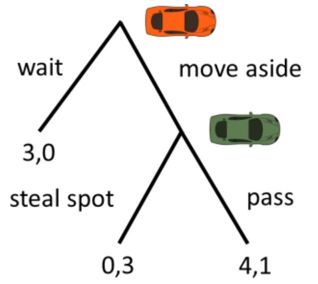
Fan gets Decent seat

Uninterested gets No seat

Summary of approach

- Game-theoretic failures to cooperate can happen **even with almost perfectly aligned agents**
- Some ways of getting to cooperation make sense for **humans** as well…
- … but there are others that seem more natural for **(advanced) AI agents**
- Let's not unnecessarily limit our toolkit!

	100	90	80	70	60	50	40	30	20	10	0
100	111, 111	90, 112	80, 102	70, 92	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
90	112, 90	101, 101	80, 102	70, 92	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
80	102, 80	102, 80	91, 91	70, 92	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
70	92, 70	92, 70	92, 70	81, 81	60, 82	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
60	82, 60	82, 60	82, 60	82, 60	71, 71	50, 72	40, 62	30, 52	20, 42	10, 32	0, 22
50	72, 50	72, 50	72, 50	72, 50	72, 50	61, 61	40, 62	30, 52	20, 42	10, 32	0, 22
40	62, 40	62, 40	62, 40	62, 40	62, 40	62, 40	51, 51	30, 52	20, 42	10, 32	0, 22
30	52, 30	52, 30	52, 30	52, 30	52, 30	52, 30	52, 30	41, 41	20, 42	10, 32	0, 22
20	42, 20	42, 20	42, 20	42, 20	42, 20	42, 20	42, 20	42, 20	31, 31	10, 32	0, 22
10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	32, 10	21, 21	0, 22
0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	22, 0	$\boxed{11, 11}$



ethics & norms

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

t=0 →

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

t=1 → …

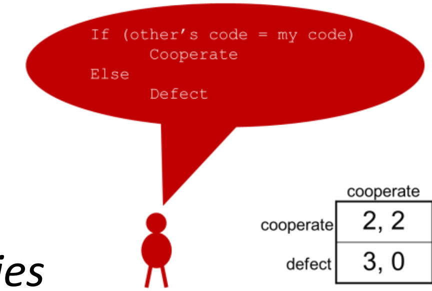
repeated games

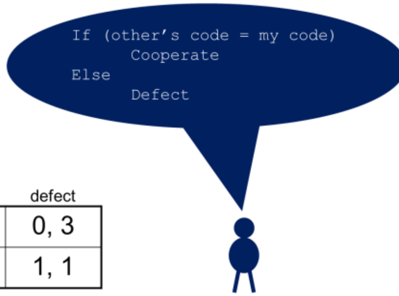
	GR		ST
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

disarmament (pure strategies)

Chair \ Member	sabotage	cooperate
selfish	1 / 2	0 / 3
cooperate	1 / 1	2 / 2

role assignment / agent boundaries





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

program equilibrium, simulation

```
1. With probability
40%, cooperate
3. With probability
40%, cooperate
...
```

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

```
2. With probability
40%, cooperate
4. With probability
40%, cooperate
...
```

disarmament (mixed strategies)

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

philosophical foundations (evidential decision theory, self-locating belief, …)

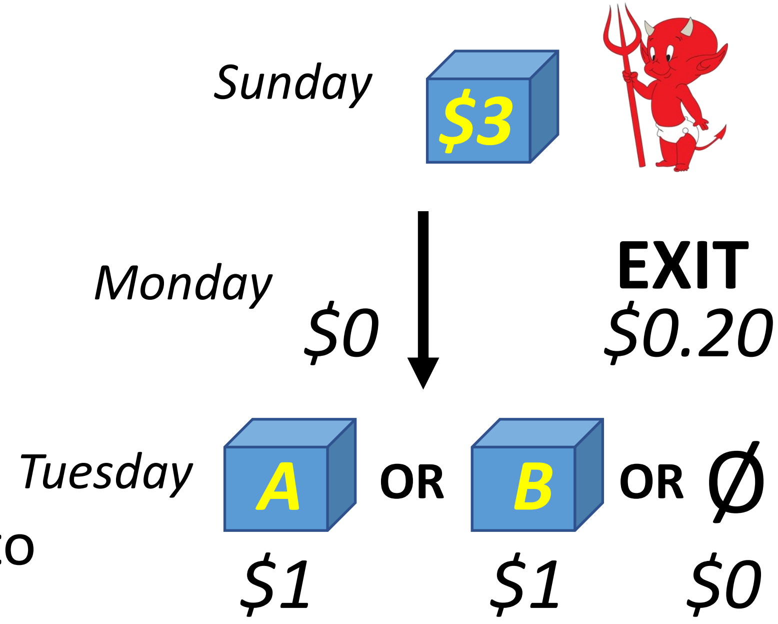
Many open questions

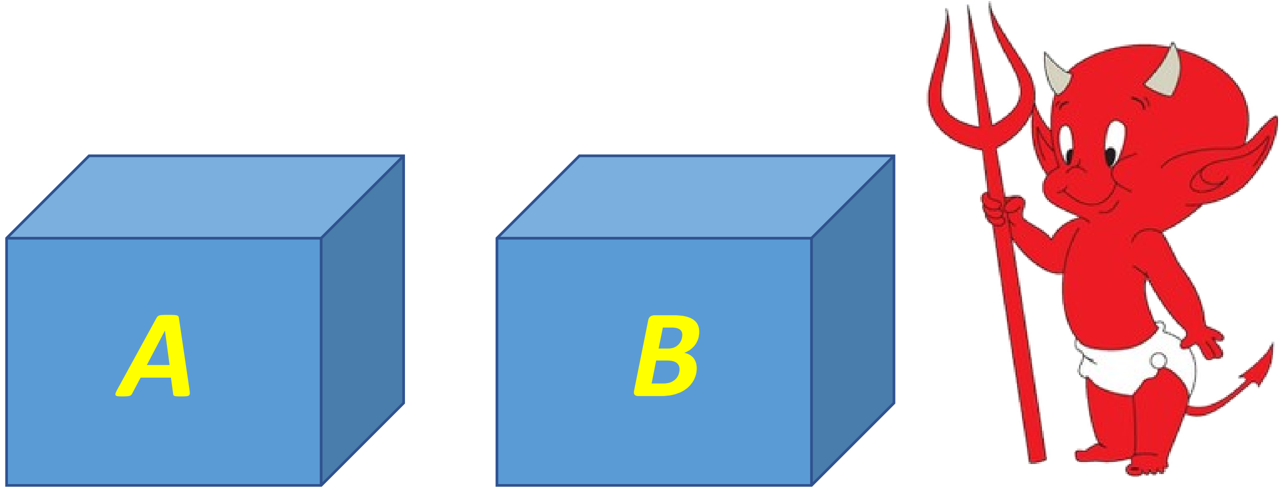
- What are the foundations of game theory for highly advanced AI?
- How should an agent play with other agents with overlapping code? With visible code?
- How should an agent play when it may be being simulated? When it can't remember the past?
- What design decisions can improve cooperation?
  - How realistic are they? How do we make them more so?
  - How robust are they? How do we make them more so?
- What is the role of learning?
  - Can we design learning algorithms that converge to good equilibria?
  - In contexts of logical uncertainty?
- …

Turning causal decision theorists into money pumps

[Oesterheld and C., Phil. Quarterly'21]

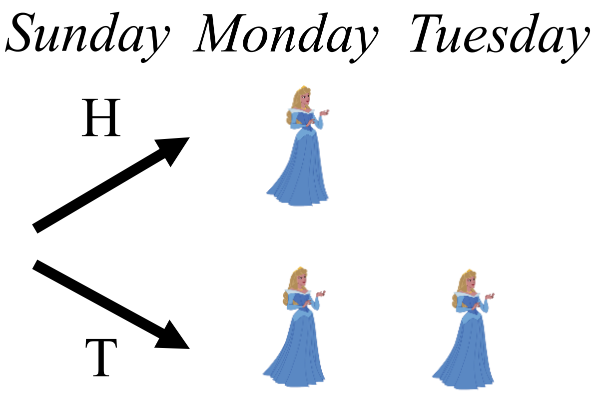
- **Adversarial Offer:**
- Demon (really, any good predictor) put $3 into each box it predicted you would not choose
- Each box costs $1 to open; can open at most one
- Demon 75% accurate (you have no access to randomization)
- CDT will choose one box, *knowing that it will regret doing so*
- Can add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





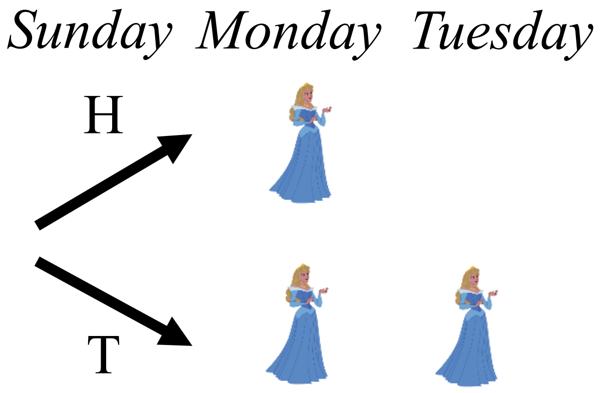
Taking advantage of a Halfer [Hitchcock'04]

- 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. 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; Oesterheld & C. working paper]
- 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

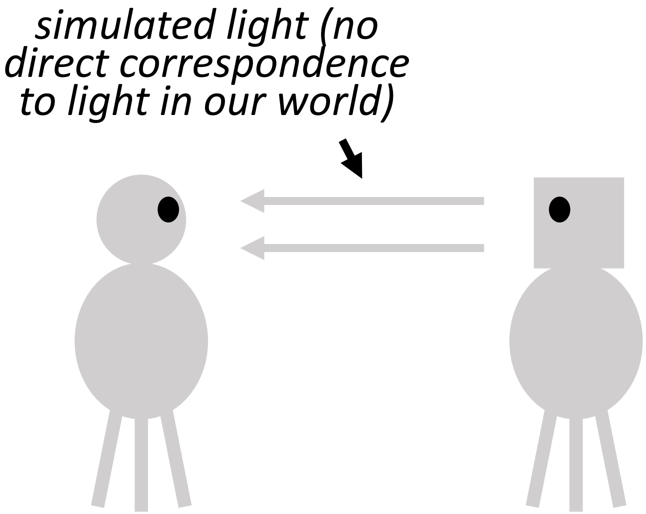
	WG (1/4)	WO (1/4)	BO (1/4)	BG (1/4)
Monday	white	white	black	black
Tuesday	grey	black	white	grey

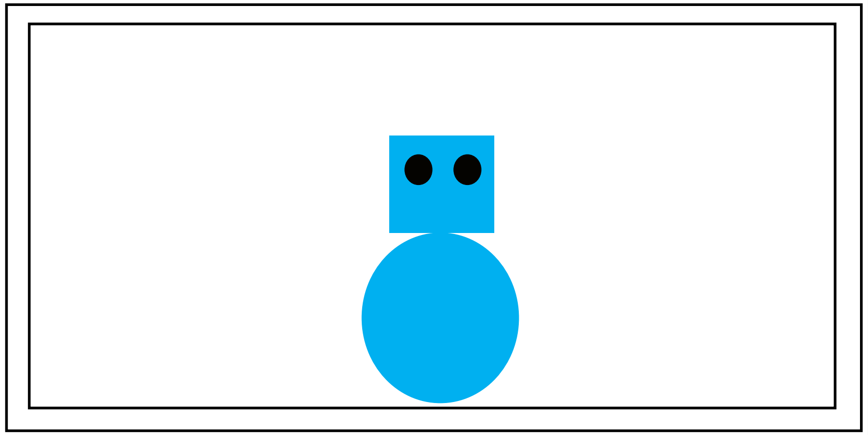
Fig. 3 Sequences of coin tosses and corresponding room colors, as well as their probabilities, in the WBG Sleeping Beauty variant.

	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.

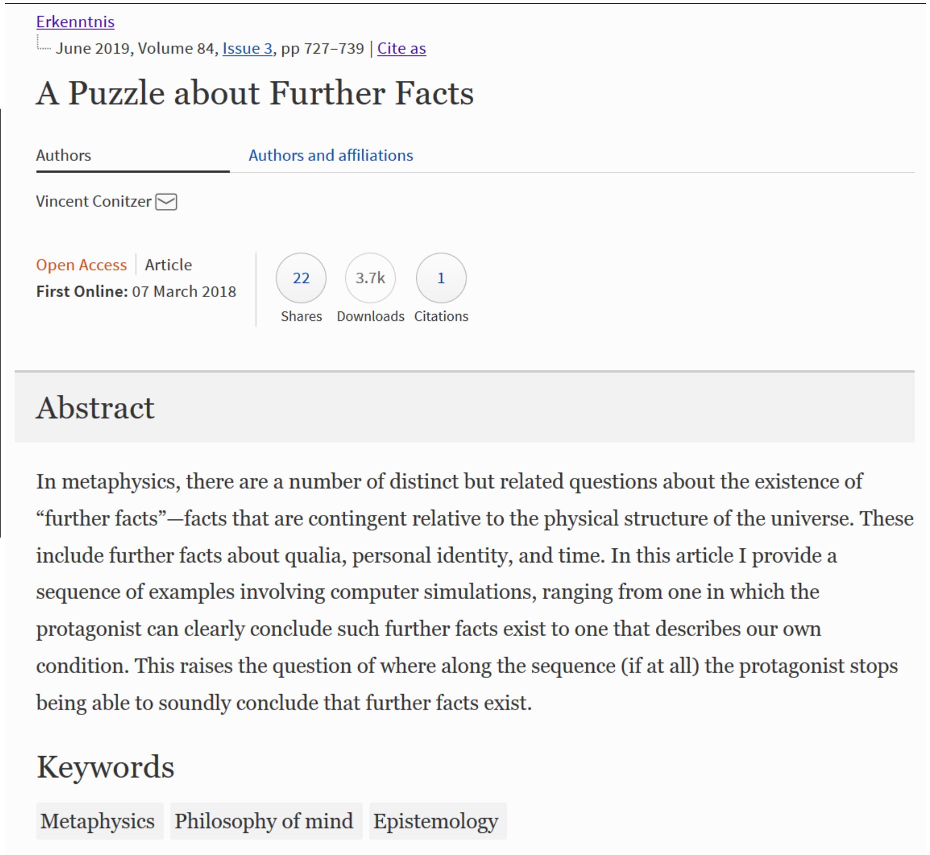
Philosophy of “being present” somewhere, sometime





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

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?
- Related to A-theory and B-theory of time in metaphysics [C., dialectica’20]

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...

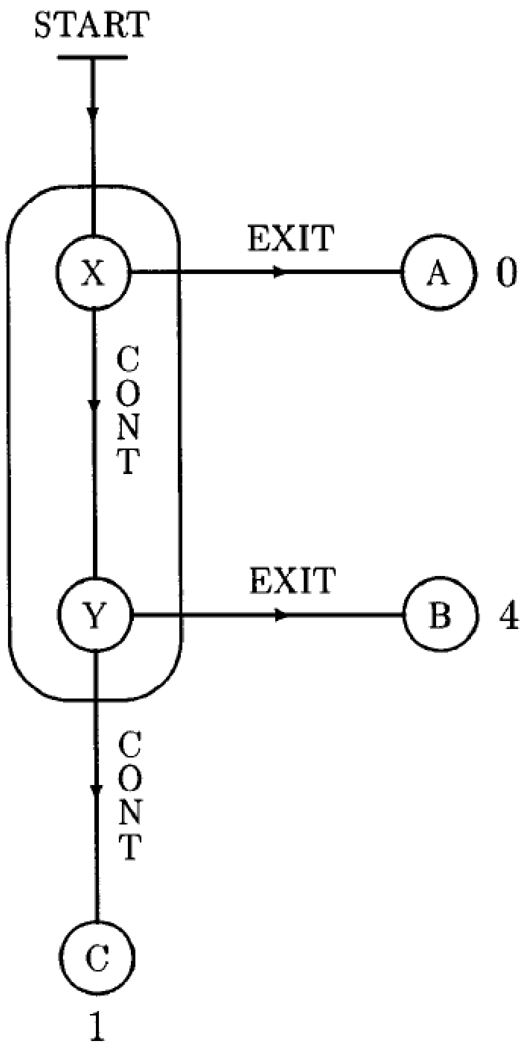


FIG. 1. The absent-minded driver problem.

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, Vincent Conitzer, and Stuart Russell. For Learning in Symmetric Teams, Local Optima are Global Nash Equilibria. ICML'22


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?

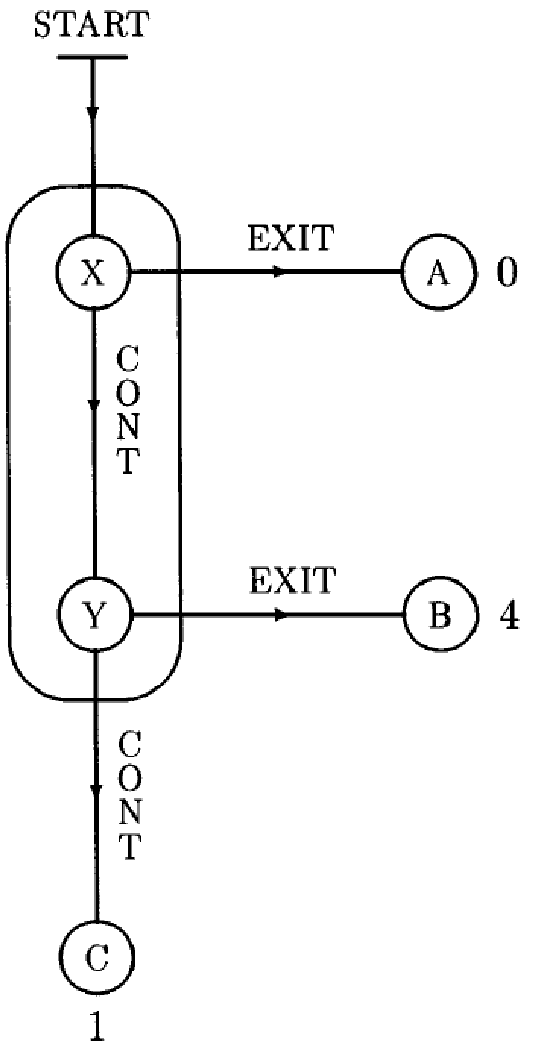
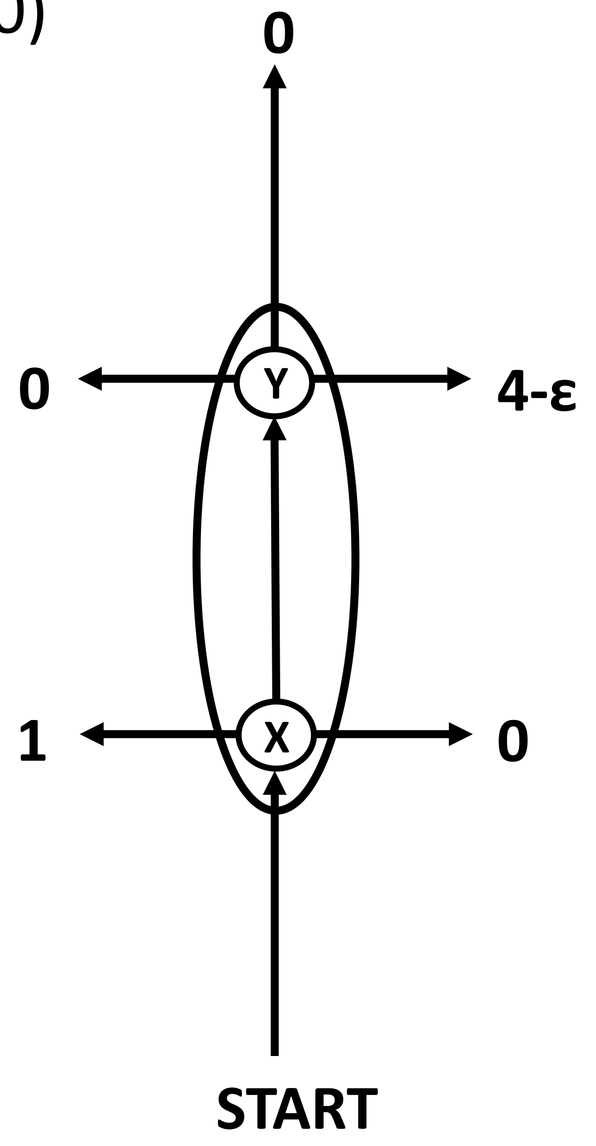


FIG. 1. The absent-minded driver problem.

Image from Aumann, Hart, Perry 1997

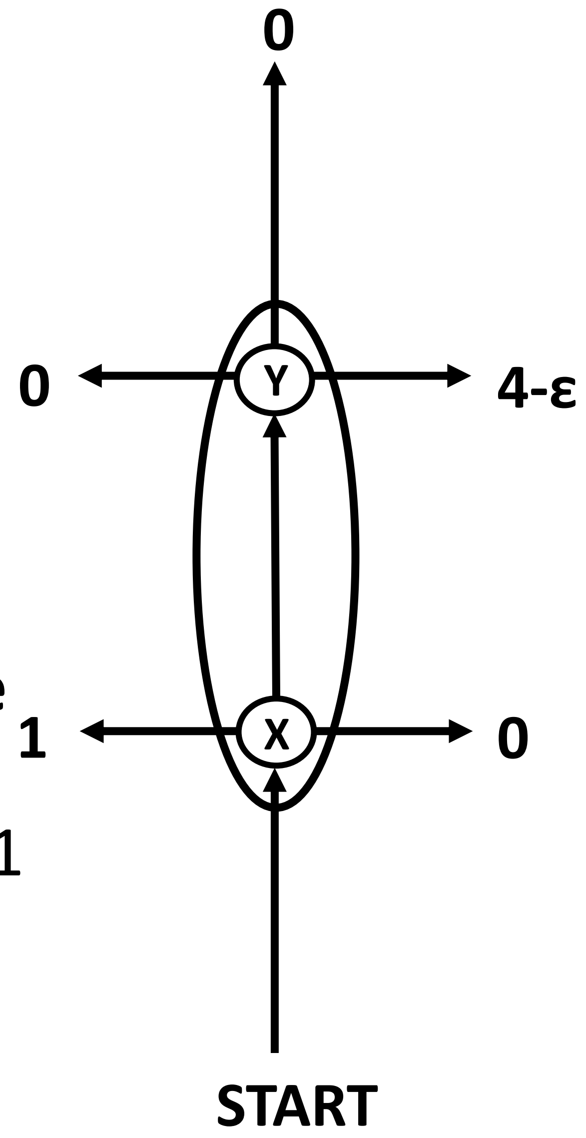
A challenging example for the evidential decision theorist

- 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)

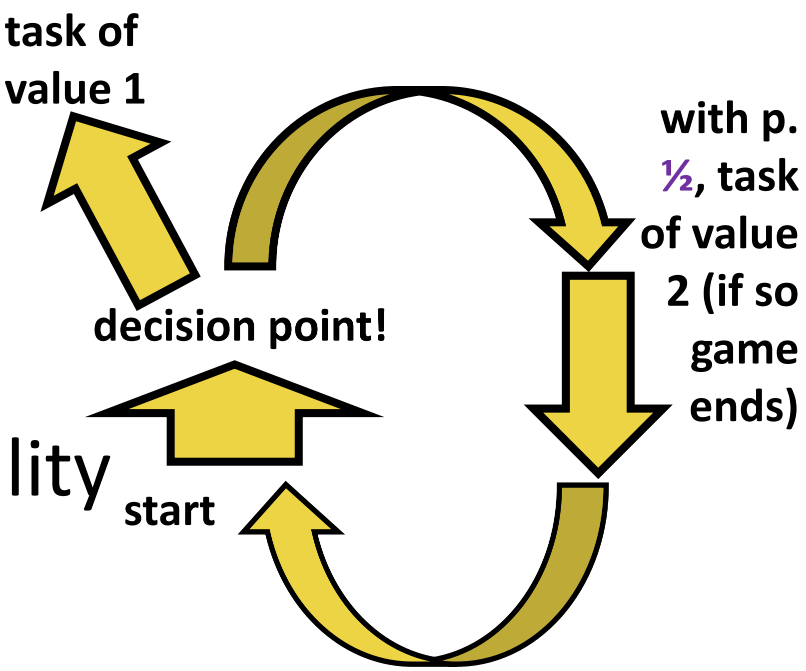
- Consider probabilities of **whole trajectories, plus where you are**, under strategy (0, ½, ½), **in a halfing sort of way**
- P(XY(4-ε), @X) = P(XY(4-ε)) * P(@X|XY(4-ε)) = ¼ * ½
- P(XY(4-ε), @Y) = P(XY(4-ε)) * P(@Y|XY(4-ε)) = ¼ * ½
- Any other trajectory with positive probability gives payoff 0
- 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?)