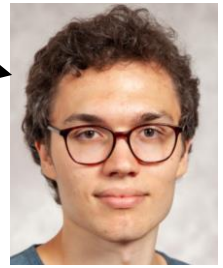
Foundations of Cooperative AI

Vincent Conitzer and
Caspar Oesterheld



Paper:
V. Conitzer and C. Oesterheld. Foundations of Cooperative AI. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*, Washington, DC, USA, 2023.

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

Outline

- Tragedies of algorithmic interaction – examples and worries
- Rethinking the design of intelligent agents
  - (Intelligence + value alignment) still allows game-theoretic tragedies
- Should AI systems cooperate like humans do?
- Techniques for achieving cooperation that (also) fit humans
- Techniques for achieving cooperation that don’t fit humans
- Open questions and call to action

Outline

- **Tragedies of algorithmic interaction – examples and worries**
- Rethinking the design of intelligent agents
  - (Intelligence + value alignment) still allows game-theoretic tragedies
- Should AI systems cooperate like humans do?
- Techniques for achieving cooperation that (also) fit humans
- Techniques for achieving cooperation that don’t fit humans
- Open questions and call to action

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.

Outline

- Tragedies of algorithmic interaction – examples and worries
- **Rethinking the design of intelligent agents**
  - **(Intelligence + value alignment) still allows game-theoretic tragedies**
- Should AI systems cooperate like humans do?
- Techniques for achieving cooperation that (also) fit humans
- Techniques for achieving cooperation that don’t fit humans
- Open questions and call to action

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

Stuart Russell

Peter Norvig

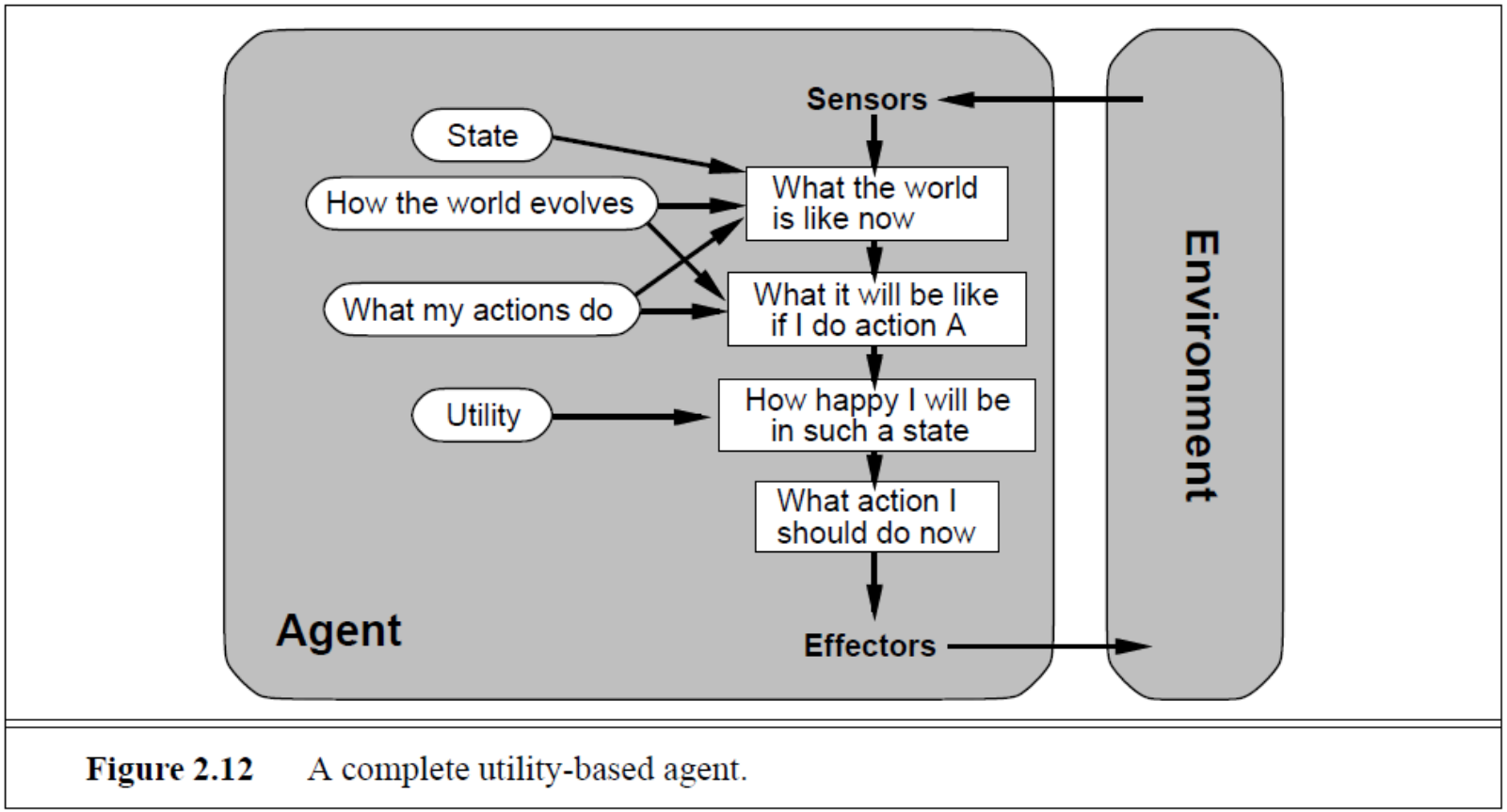
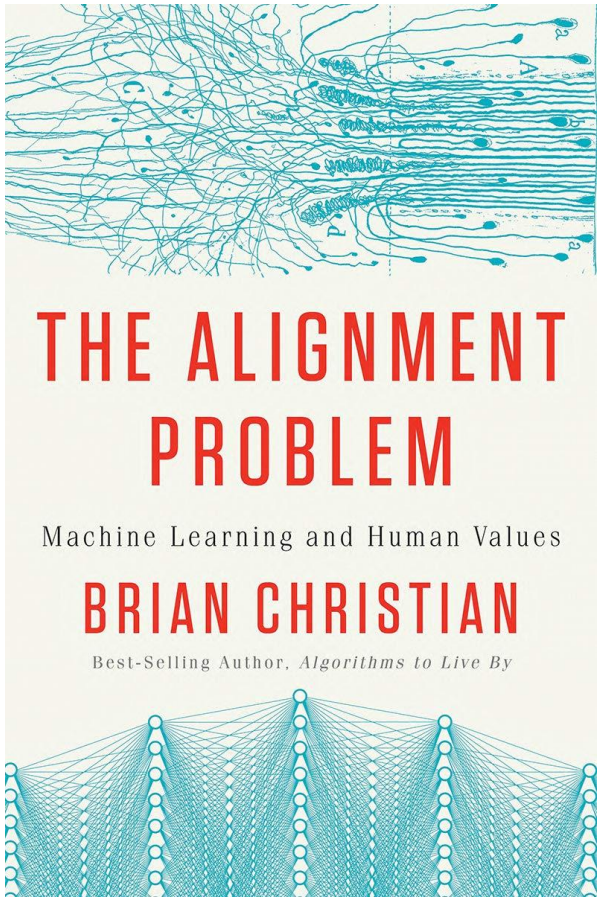
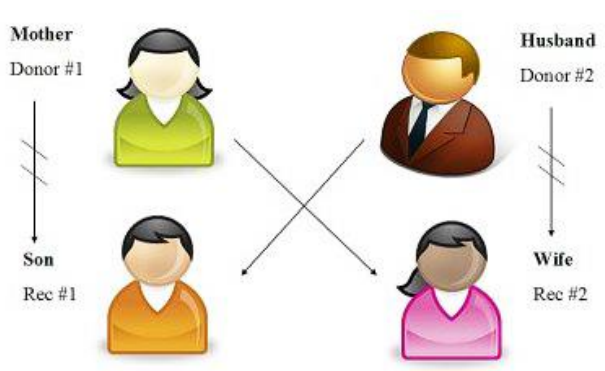


Figure 2.12 A complete utility-based agent.

"… 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

FAccT 서울 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

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

Outline

- Tragedies of algorithmic interaction – examples and worries
- Rethinking the design of intelligent agents
  - (Intelligence + value alignment) still allows game-theoretic tragedies
- **Should AI systems cooperate like humans do?**
- Techniques for achieving cooperation that (also) fit humans
- Techniques for achieving cooperation that don’t fit humans
- Open questions and call to action

Science

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

NEWS | BRAIN & BEHAVIOR

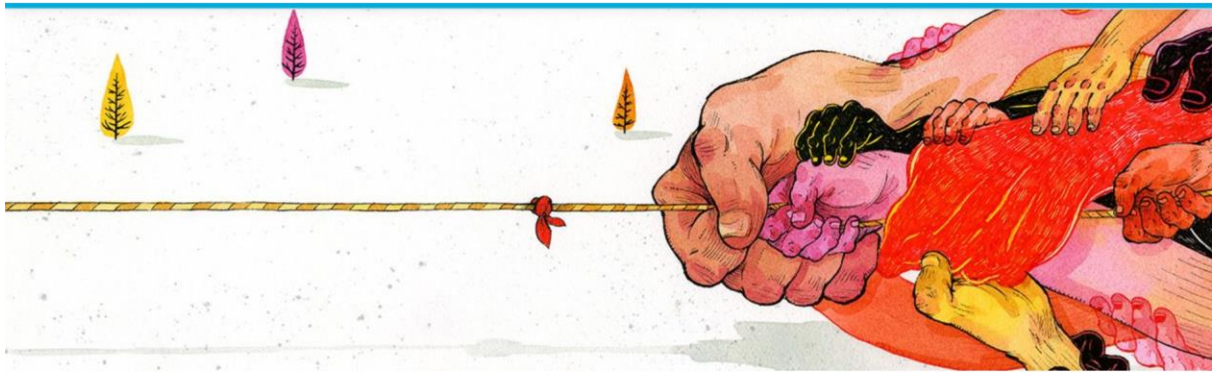
Human altruism traces back to the origins of humanity

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. 2010 Sep 12; 365(1553): 2663–2674.
doi: 10.1098/rstb.2010.0157

PMCID: PMC2936178
PMID: 20679110

Philos Trans R Soc Lond B Biol Sci

How is human cooperation different?

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

▸ Author information ▸ Copyright and License information Disclaimer

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

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

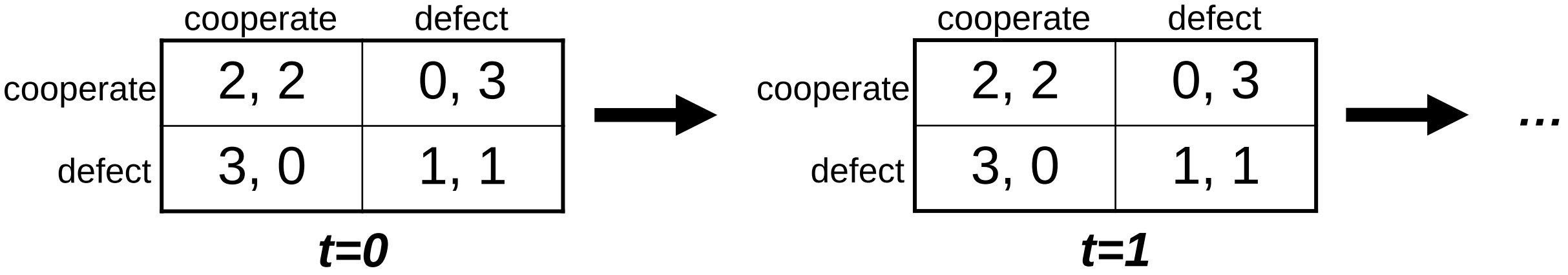
Posted on December 15, 2019, at 10:29 a.m. ET

Tweet Share Copy

Outline

- Tragedies of algorithmic interaction – examples and worries
- Rethinking the design of intelligent agents
  - (Intelligence + value alignment) still allows game-theoretic tragedies
- Should AI systems cooperate like humans do?
- **Techniques for achieving cooperation that (also) fit humans**
- Techniques for achieving cooperation that don’t fit humans
- Open questions and call to action

Infinitely Repeated Prisoner's Dilemma



- **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

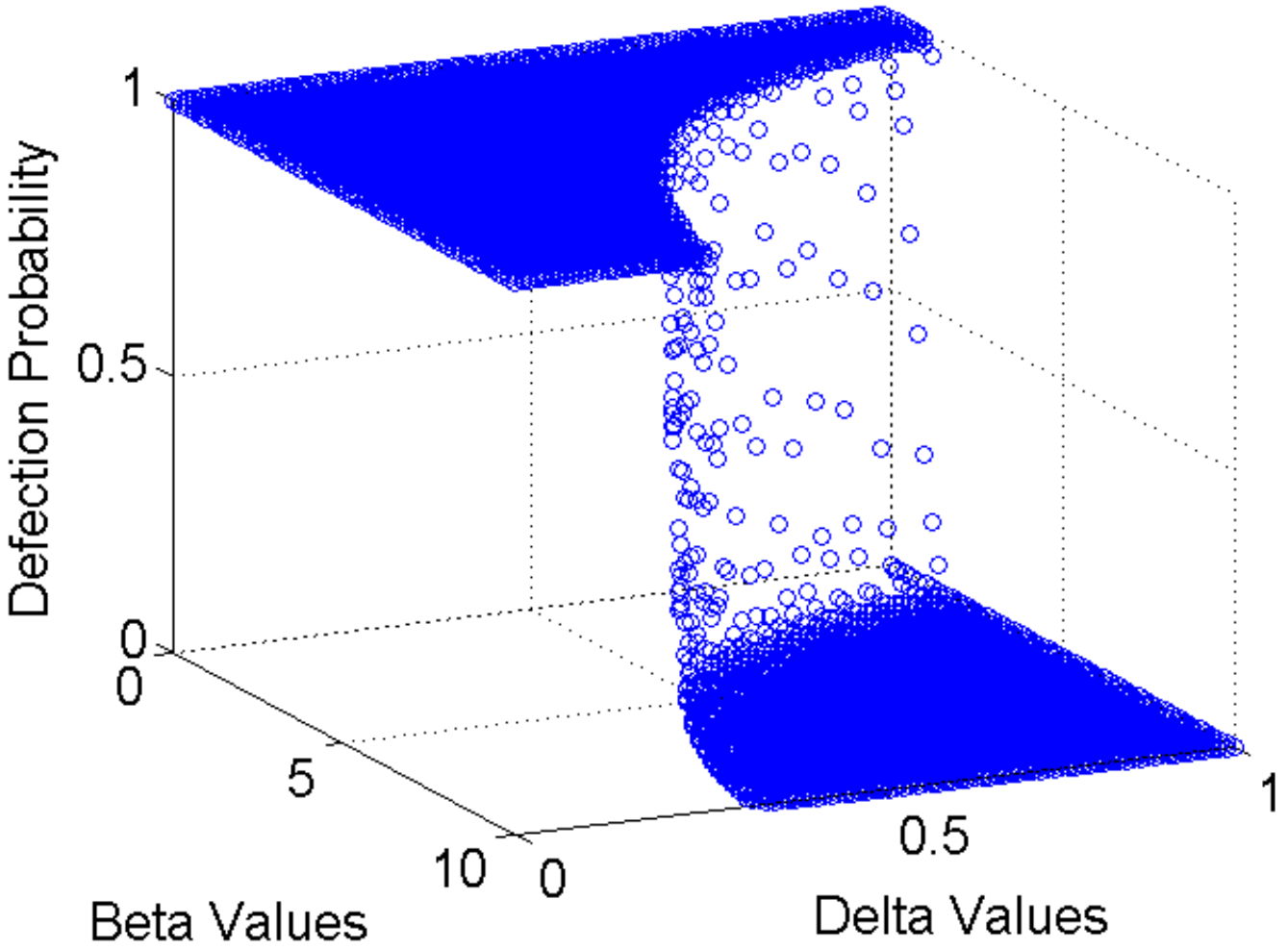
[Moon & C., IJCAI'15]

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

Outline

- Tragedies of algorithmic interaction – examples and worries
- Rethinking the design of intelligent agents
  - (Intelligence + value alignment) still allows game-theoretic tragedies
- Should AI systems cooperate like humans do?
- Techniques for achieving cooperation that (also) fit humans
- **Techniques for achieving cooperation that don’t fit humans**
- Open questions and call to action

Program equilibrium [Tennenholz 2004]

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

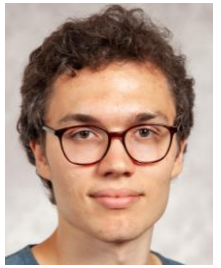
```
If (other's code = my code)
      Cooperate
Else
      Defect
```

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

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

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

Robust program equilibrium [Oesterheld 2018]



Caspar Oesterheld

- Can we make the equilibrium less fragile?

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

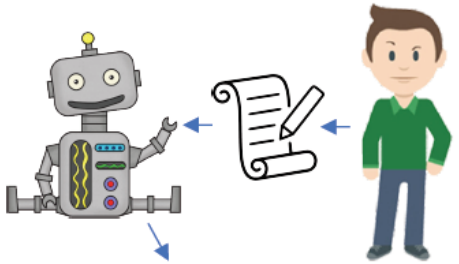
…

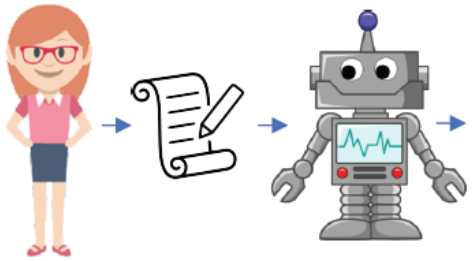
	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

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

Caspar Oesterheld

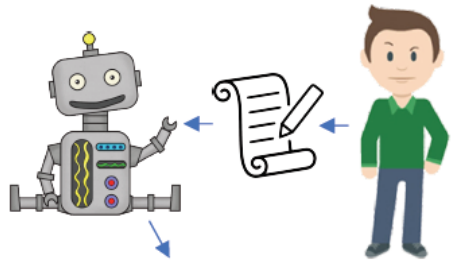
Delegated game playing

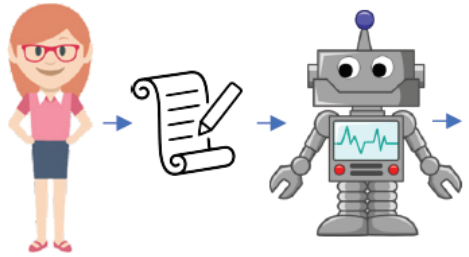




	DM	RM	DL	RL
DM	-5,-5	2,0	5,-5	5,-5
RM	0,2	1,1	5,-5	5,-5
DL	-5,5	-5,5	1,1	2,0
RL	-5,5	-5,5	0,2	1,1

- 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

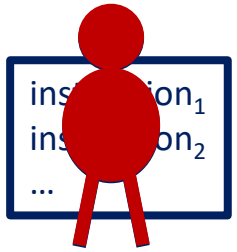


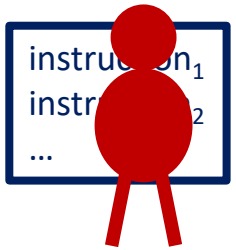


	DL	RL
DL	-5,-5 (1,1)	2,0 (2,0)
RL	0,2 (0,2)	1,1 (1,1)

- 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.

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?

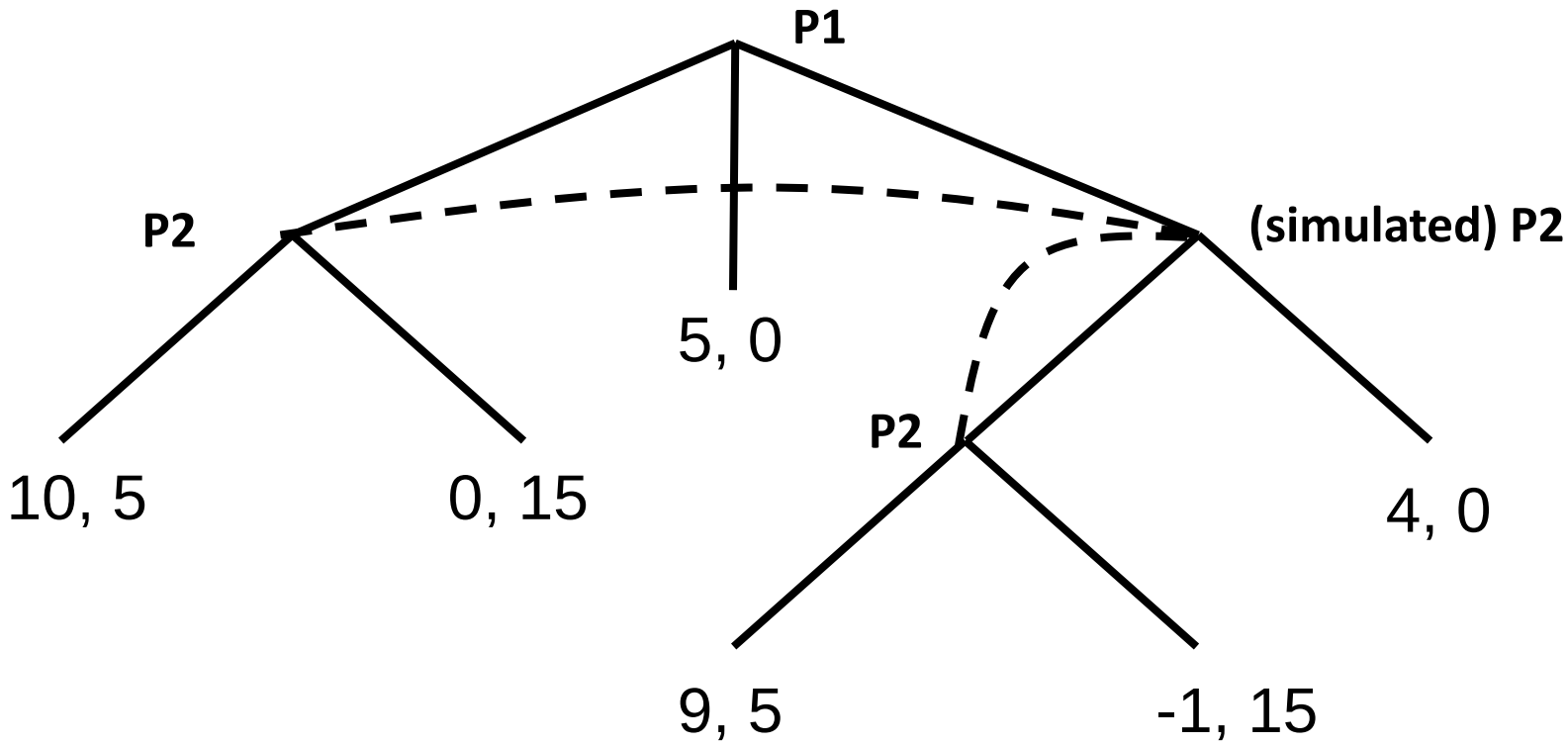
related to working paper
[Oesterheld, Demski, C.]

Caspar Oesterheld

Abram Demski

Simulating our way to cooperation?

- Restricted trust game: 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

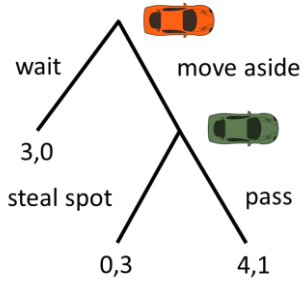


As (AI system) P2, how likely is it you're now running *as a simulation*? → *self-locating belief*

What happens in equilibrium?

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!



ethics & norms

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

t=0 → *t=1* → …

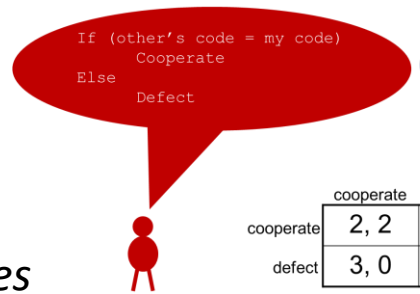
repeated games

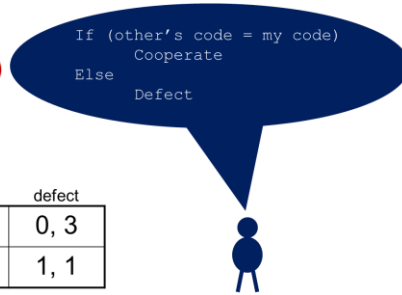
	GR		ST
PF	4, 1		-0.5, -0.5
PD	-6, 8		-0.5, -0.5

disarmament (pure strategies)

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

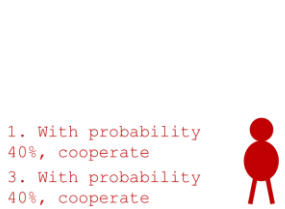
role assignment / agent boundaries



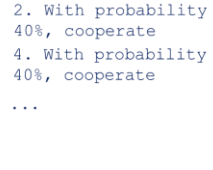


	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

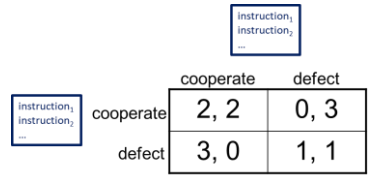
program equilibrium



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



disarmament (mixed strategies)



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

Outline

- Tragedies of algorithmic interaction – examples and worries
- Rethinking the design of intelligent agents
  - (Intelligence + value alignment) still allows game-theoretic tragedies
- Should AI systems cooperate like humans do?
- Techniques for achieving cooperation that (also) fit humans
- Techniques for achieving cooperation that don’t fit humans
- **Open questions and call to action**

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?
- …

THANK YOU FOR
YOUR ATTENTION!