TRAIL & VAIA

AI seminars

Cooperative AI

Game theoretical research in socially beneficial AI

Tom Lenaerts 18/11/2022

tomlenaerts.be

MLG MACHINE LEARNING GROUP

ULB

ARTIFICIAL INTELLIGENCE RESEARCH GROUP

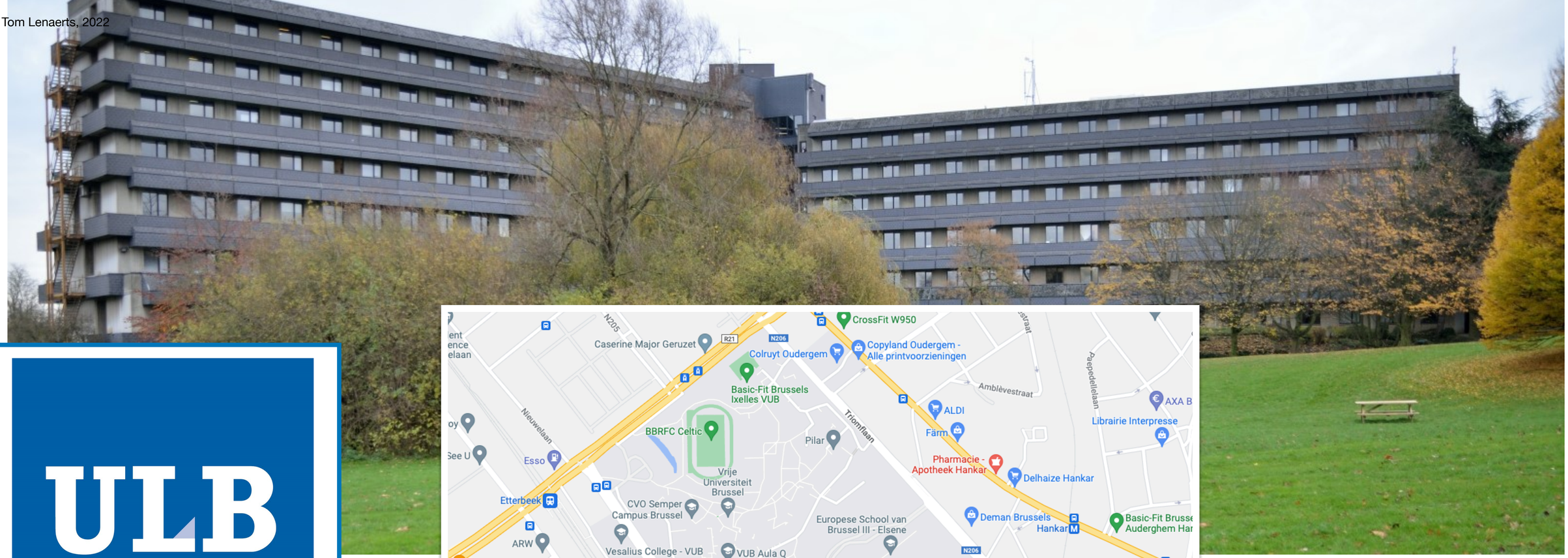
VUB

FARI

AI FOR THE COMMON GOOD INSTITUTE BRUSSELS

CHAI Center for Human-Compatible Artificial Intelligence

B2 INTERUNIVERSITY INSTITUTE OF BIOINFORMATICS IN BRUSSELS



ULB

MLG MACHINE LEARNING GROUP

Département d'Informatique, Faculté des Sciences, Université Libre de Bruxelles, Boulevard du Triomphe CP212. 1050 Brussels, Belgium

di.ulb.ac.be

ULB created in 1834

~1000 CS students (Science faculty)

> **4 research groups** (17 professors and their teams)

Created in 2004 by **Gianluca Bontempi**

Co-headed by Gianluca and Tom since 2010

ULB

MLG
MACHINE LEARNING GROUP

~24 researchers (4 professors, ~18 PhDs and 2 postdocs)

> 450 publications, covering a wide range of ML, AI, optimisation, statistics and domain-specific topics (e.g. medical/biological, mobility, fraud, …)

>11 ongoing projects (EU, Future of Life institute, NESTA collective intelligence, FWO, FNRS, Innoviris, DigitalWallonia.ai…)

http://mlg.ulb.ac.be

mlgulb

MLG-ulb

firstname.lastname@ulb.be

TRAIL & VUB AI seminars

Time series analysis

Feature selection

Interpretability

Machine learning & Big data analysis

ULB

MLG MACHINE LEARNING GROUP

Artificial intelligence

Evolutionary dynamics

Bioinformatics and Computational biology

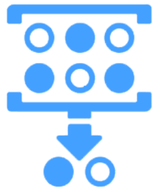
Game theory

AI governance

Collective Intelligence

Time series analysis

Feature selection



Interpretability

Machine learning &
Big data analysis

ULB VUB

INTERUNIVERSITY INSTITUTE OF BIOINFORMATICS IN BRUSSELS

http://www.ibsquare.be

Bioinformatics and
Computational biology

ULB

MLG
MACHINE LEARNING GROUP

Artificial intelligence

Evolutionary dynamics

Game theory

AI governance

Collective
Intelligence



ULB VUB

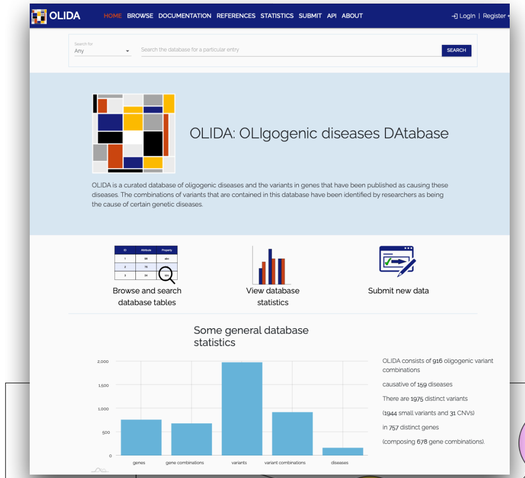
Interuniversity Institute of Bioinformatics in Brussels

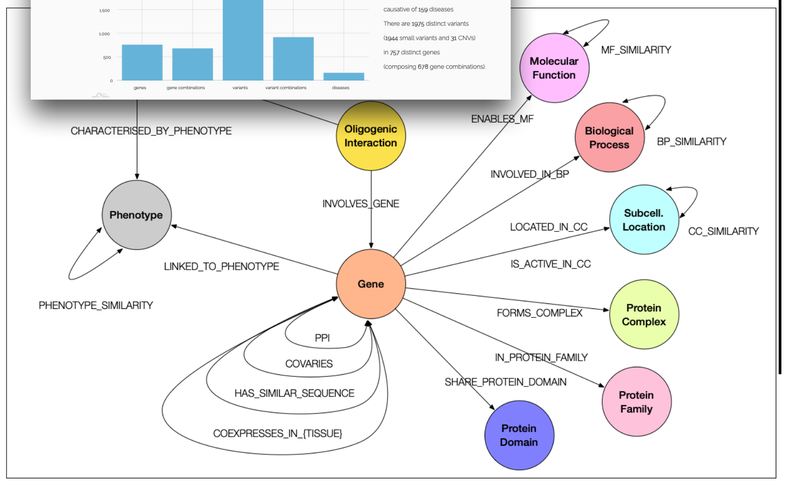
http://www.ibsquare.be




Biocuration/Active learning for text mining/Knowledge graphs/embeddings

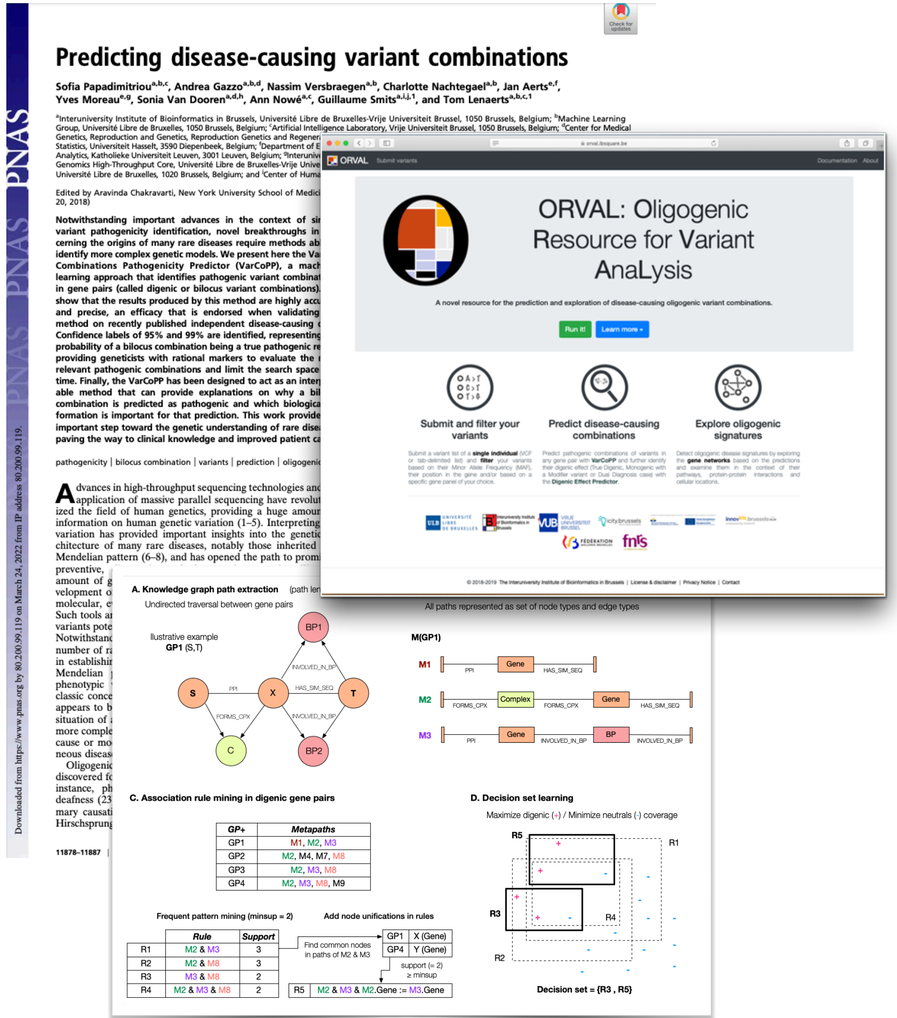
http://olida.ibsquare.be





Oligogenic knowledge graph

Precision Medicine (ML/rule mining)



Predicting disease-causing variant combinations

Sofia Papadimitriou[a,b,c], Andrea Gazzo[a,b,d], Nassim Versbraegen[a,b], Charlotte Nachtegael[a,b], Jan Aerts[e,f], Yves Moreau[e,g], Sonia Van Dooren[a,d,h], Ann Nowé[a,c], Guillaume Smits[a,i,j,1], and Tom Lenaerts[a,b,c,1]

[a]Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussel, 1050 Brussels, Belgium; [b]Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium; [c]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, 1050 Brussels, Belgium; [d]Center for Medical Genetics, Reproduction and Genetics, Reproduction Genetics and Regener… Statistics, Universiteit Hasselt, 3590 Diepenbeek, Belgium; [f]Department of E… Analytics, Katholieke Universiteit Leuven, 3001 Leuven, Belgium; [g]Interuniv… Genomics High-Throughput Core, Université Libre de Bruxelles-Vrije Unive… Université Libre de Bruxelles, 1020 Brussels, Belgium; and [h]Center of Huma…

Edited by Aravinda Chakravarti, New York University School of Medici… 20, 2018)

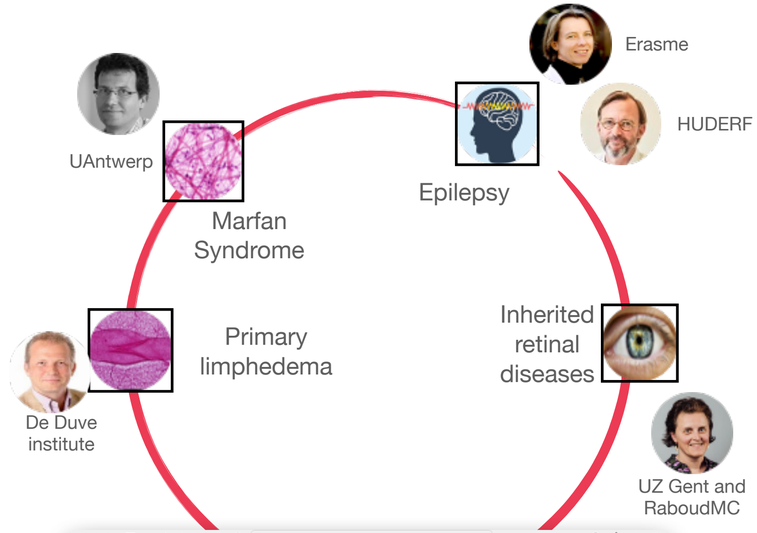
Notwithstanding important advances in the context of si… variant pathogenicity identification, novel breakthroughs in … cerning the origins of many rare diseases require methods ab… identify more complex genetic models. We present here the Va… Combinations Pathogenicity Predictor (VarCoPP), a mach… learning approach that identifies pathogenic variant combinat… in gene pairs (called digenic or bilocus variant combinations). … show that the results produced by this method are highly accu… and precise, an efficacy that is endorsed when validating … method on recently published independent disease-causing … Confidence labels of 95% and 99% are identified, representing … probability of a bilocus combination being a true pathogenic re… providing geneticists with rational markers to evaluate the … relevant pathogenic combinations and limit the search space … time. Finally, the VarCoPP has been designed to act as an inter… able method that can provide explanations on why a bi… combination is predicted as pathogenic and which biologica… formation is important for that prediction. This work provide… important step toward the genetic understanding of rare dise… paving the way to clinical knowledge and improved patient ca…

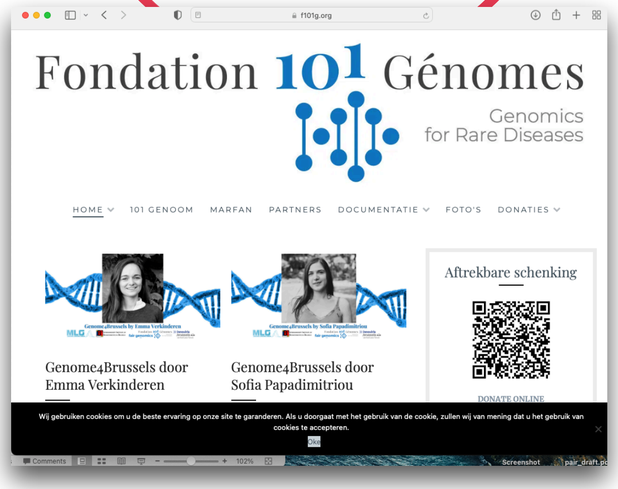
pathogenicity | bilocus combination | variants | prediction | oligogeni…

Advances in high-throughput sequencing technologies and … application of massive parallel sequencing have revolut… ized the field of human genetics, providing a huge amoun… information on human genetic variation (1–5). Interpreting … variation has provided important insights into the genetic … chitecture of many rare diseases, notably those inherited … Mendelian pattern (6–8), and has opened the path to promi… preventive, … amount of g… velopment o… molecular, e… Such tools a… variants pote… Notwithstand… number of ri… in establishi… Mendelian … phenotypic … classic conce… appears to b… situation of … more comple… cause or mo… neous disea… Oligogenic … discovered fo… instance, ph… deafness (23… mary causati… Hirschsprun…

11878–11887 | …

Validation/clinics/patients





Time series analysis

Machine learning &
Big data analysis

Interpretability

Feature selection

ULB

MLG
MACHINE LEARNING GROUP

Artificial intelligence

Evolutionary dynamics

Game theory

Bioinformatics and
Computational biology

AI governance

Collective
Intelligence

Setting the agenda in research

Comment

A huddle at the 2017 United Nations Climate Change Conference, where attendees cooperated on mutually beneficial joint actions on climate.

Cooperative AI: machines must learn to find common ground

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel

To help humanity solve fundamental problems of cooperation, scientists need to reconceive artificial intelligence as deeply social.

Artificial-intelligence assistants and recommendation algorithms interact with billions of people every day, influencing lives in myriad ways, yet they still have little understanding of humans. Self-driving vehicles controlled by artificial intelligence (AI) are gaining mastery of their interactions with the natural world, or collaborating with their human operators. The state of AI applications reflects that of the research field. It has long been steeped in a kind of methodological individualism. As is evident from introductory textbooks, the canonical AI problem is that of a solitary machine confronting a non-social environment. Historically, this was a sensible start-

The problem of cooperation between AI and/or humans

*"The coming years might give rise to diverse ecologies of AI systems that interact in rapid and complex ways with each other and with humans … Autonomous vehicles and smart cities that do not engage well with humans will fail to deliver … **we need to build a science of cooperative AI**"*

Not a new question, but …

Panait, L., & Luke, S. (2005). Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, *11*(3), 387-434.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994* (pp. 157-163). Morgan Kaufmann.

Doran, J. E., Franklin, S. R. J. N., Jennings, N. R., & Norman, T. J. (1997). On cooperation in multi-agent systems. *The Knowledge Engineering Review*, *12*(3), 309-314.

Vittikh, V. A., & Skobelev, P. O. (**1970**). Multi-agent systems for modelling of self-organization and cooperation processes. *WIT Transactions on Information and Communication Technologies*, *20*.

Setting the agenda in research

Comment

SEAN GALLUP/GETTY

A huddle at the 2017 United Nations Climate Change Conference, where attendees cooperated on mutually beneficial joint actions on climate.

Cooperative AI: machines must learn to find common ground

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel

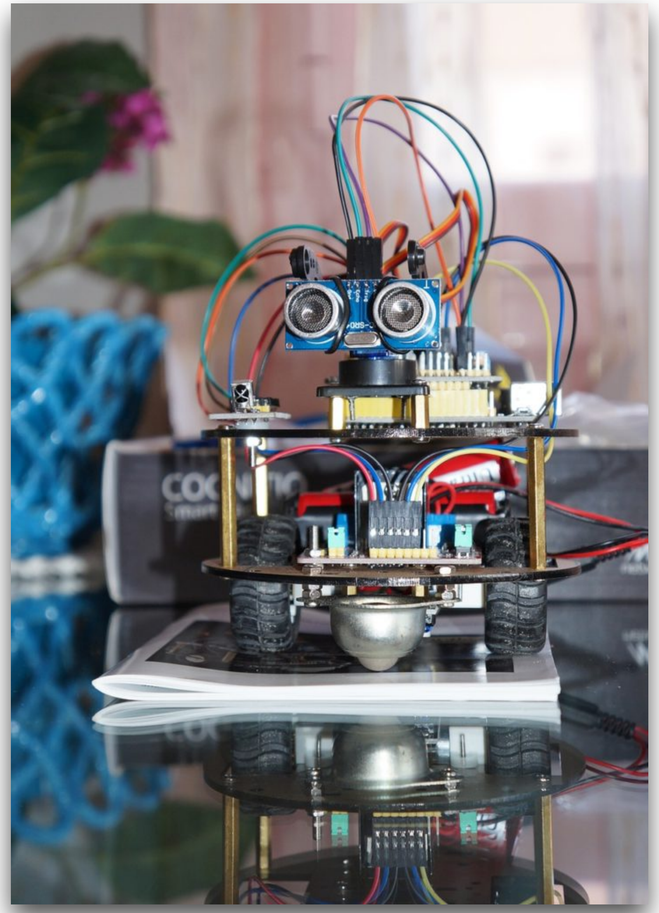
To help humanity solve fundamental problems of cooperation, scientists need to reconceive artificial intelligence as deeply social.

Artificial-intelligence assistants and recommendation algorithms interact with billions of people every day, influencing lives in myriad ways, yet they still have little understanding of humans. Self-driving vehicles controlled by artificial intelligence (AI) are gaining mastery of their interactions with the natural world, or collaborating with their human operators.

The state of AI applications reflects that of the research field. It has long been steeped in a kind of methodological individualism. As is evident from introductory textbooks, the canonical AI problem is that of a solitary machine confronting a non-social environment. Historically, this was a sensible start-

Dafoe, Allan, et al. "Cooperative AI: machines must learn to find common ground." (2021): Nature 33-36.

Real world complexities and …



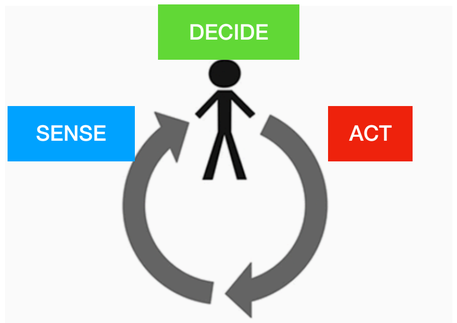
… the problem of the idiot savant in a vacuum

The problem of autonomy

How much error can we tolerate?

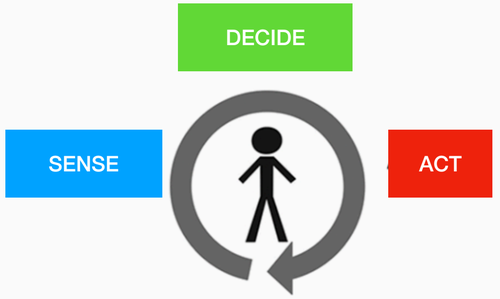
Do human and machine objectives/solutions align?

Full control



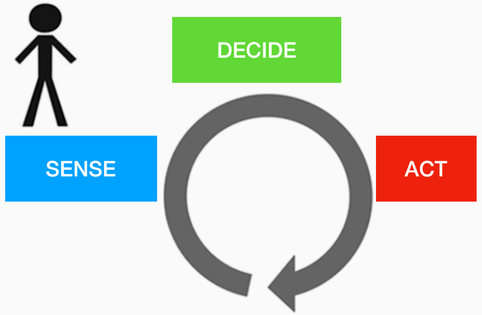
The machine performs a task and then waits for the human user to take an action before continuing.

Partial control



The machine can sense, decide, and act on its own. The human user supervises its operation and can intervene, if desired.

Full autonomy

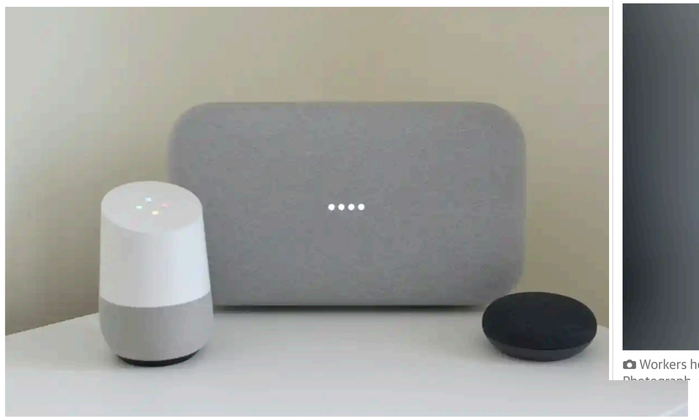


The machine can sense, decide, and act on its own. The human cannot intervene in a timely fashion.

The problem of multiple stakeholders

Google workers can listen to what people say to its AI home devices

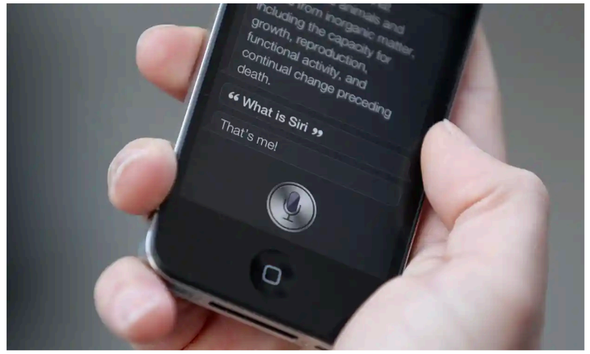
Company admitted that contractors can access recordings made by Assistant, after some of its recordings were leaked



In 2017, Google confirmed a bug in its Home Mini speaker allowed the smart device to record users even when it was not activated by the wake-up word. Photograph: Samuel Gibbs/The Guardian

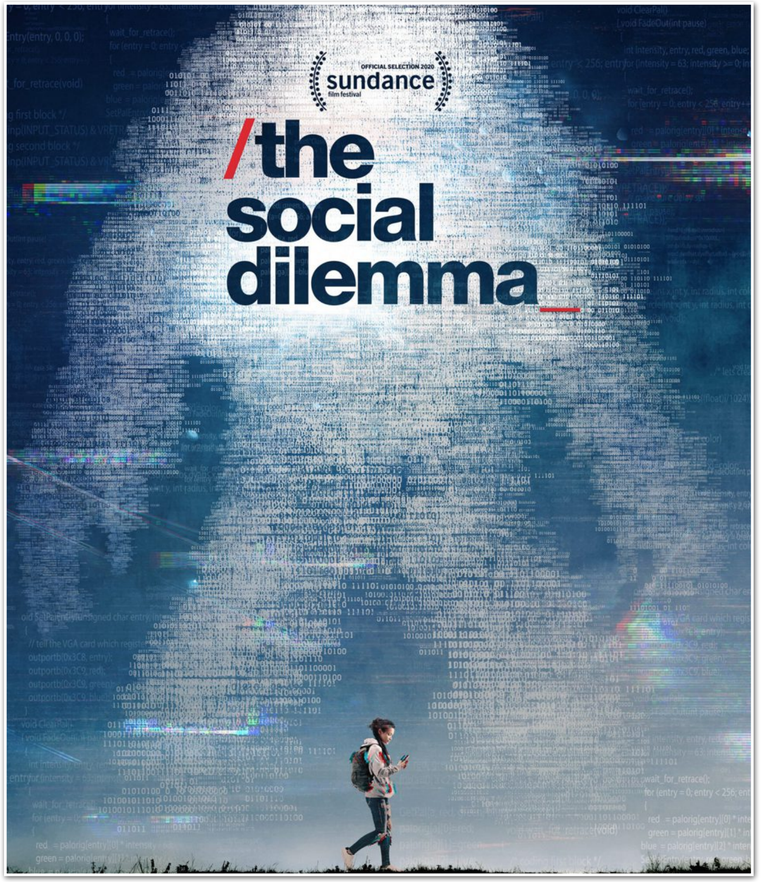
Apple contractors 'regularly hear confidential details' on Siri recordings

Workers hear drug deals, medical details and people having sex, says whistleblower



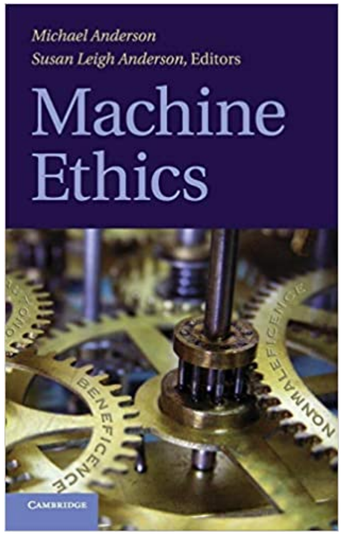
Workers heard the information when or providing quality control for Apple's Siri voice assistant.

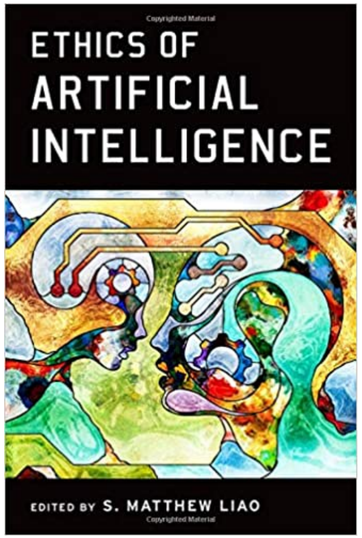
The goals of an “AI” (*and its creators*) may not be aligned with yours



Netflix theatrical poster

Governance …





And many more …

How can we **avoid** that AI's are used **that violate our norms**

How to ensure that **society as a whole benefits** from AI developments

How to **regulate** Ai developments **to avoid disasters,** harming society and its individuals

Collingridge Dilemma

"Efforts to influence or control the further development of technology face a double bind problem"

An information problem : impacts cannot be easily predicted until the technology is extensively developed and widely used

Power problem : control and change is difficult when the technology becomes entrenched

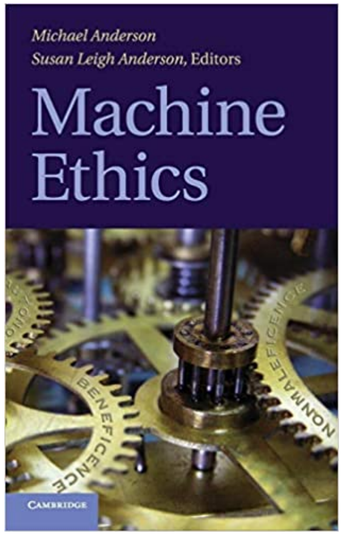
In order to understand AI governance, dynamic systems models are needed.

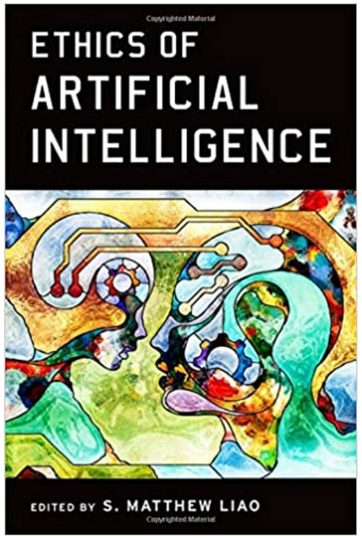
Journal of Artificial Intelligence Research 69 (2020) 881-921 Submitted 06/2020; published 11/2020

To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race

The Anh Han T.HAN@TEES.AC.UK
School of Computing, Engineering and Digital Technologies,

Governance …





And many more …

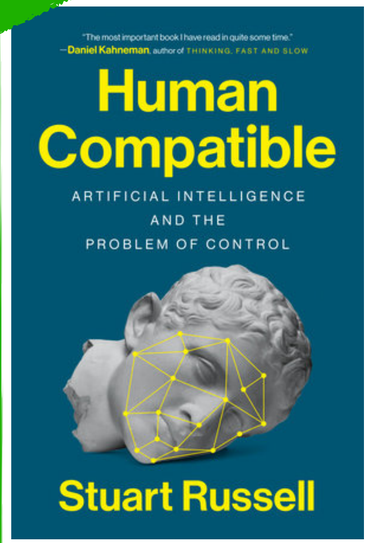
How can we **avoid** that AI's are used **that violate our norms**

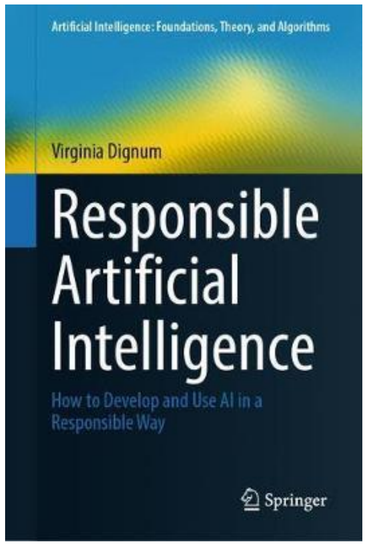
How to ensure that **society as a whole benefits** from AI developments

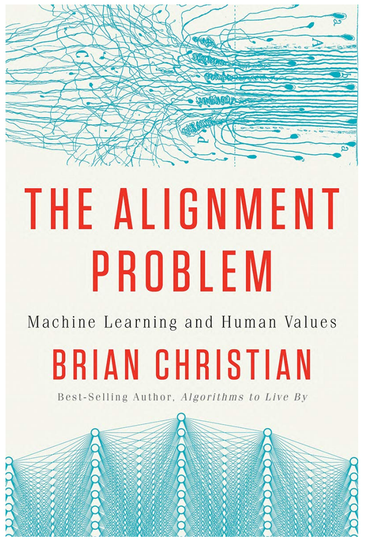
How to **regulate** Ai developments **to avoid disasters,** harming society and its individuals

Our focus

Design …







And many more …

How to **adapt the rational AI paradigm** to meet these concerns?

How to **(inter)act according to human/societal preferences** and norms?

Avoid technology solutionism !!

4 elements of cooperative intelligence need to be realised:

Understanding

AI needs a theory of mind, both affective and cognitive,

Communication

Credibly and explicitly share information,

Commitment

Have the capacity to uphold promises and

Norms and institutions

Needs social supervision so that shared beliefs and rules are followed

"To succeed, cooperative AI must connect with the broader science of cooperation, which ***spans social, behavioral and natural sciences"***

Setting the agenda in research

Comment

SEAN GALLUP/GETTY

A huddle at the 2017 United Nations Climate Change Conference, where attendees cooperated on mutually beneficial joint actions on climate.

Cooperative AI: machines must learn to find common ground

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel

To help humanity solve fundamental problems of cooperation, scientists need to reconceive artificial intelligence as deeply social.

Artificial-intelligence assistants and recommendation algorithms interact with billions of people every day, influencing lives in myriad ways, yet they still have little understanding of humans. Self-driving vehicles controlled by artificial intelligence (AI) are gaining mastery of their interactions with the natural world, or collaborating with their human operators.

The state of AI applications reflects that of the research field. It has long been steeped in a kind of methodological individualism. As is evident from introductory textbooks, the canonical AI problem is that of a solitary machine confronting a non-social environment. Historically, this was a sensible start-

Introducing game theory

YouTube video starting at 4:12

“Golden Balls is a British daytime game show which was presented by Jasper Carrott. It was broadcast on the ITV network from 18 June 2007 to 18 December 2009. It was filmed at the BBC Television Centre. Golden Balls Ltd licensed their name to Endemol for the game show and merchandise.” [Wikipedia Oct. 2020]

Sarah and Steve playing the golden balls game for 100150 pound

Players


Sarah


Steve

Actions ∈ {split, steal}

Preferences over actions:

Both prefer 100150, over 50075, over 0


(steal, split) > (split, split) > (split, steal)=(split,split)


(steal, split) > (split, split) > (split,steal)=(split,split)

We call this a **symmetric** game

Normal form of the game

		Sarah: SPLIT	Sarah: STEAL
Steve: SPLIT		50075£	100150£
		50075£	0£
Steve: STEAL		0£	0£
		100150£	0£


Steve

The simultaneous choice of both players is a **strategy profile**, e.g. (Split, Steal)

Sarah and Steve playing the golden balls game for 100150 pound

The **Nash** equilibrium

A social norm: if everyone follows it, no person will wish to deviate from this

Normal form of the game

Steve \ Sarah	SPLIT	STEAL
SPLIT	50075£, 50075£	100150£, 0£
STEAL	0£, 100150£	0£, 0£

Sarah and Steve playing the golden balls game for 100150 pound

Finding the **Nash** equilibrium

The combination of actions of the players a^* (*strategy profile*) is a Nash equilibrium if and only if every player's *i* action is a **best response** (B_i) to the other player's action

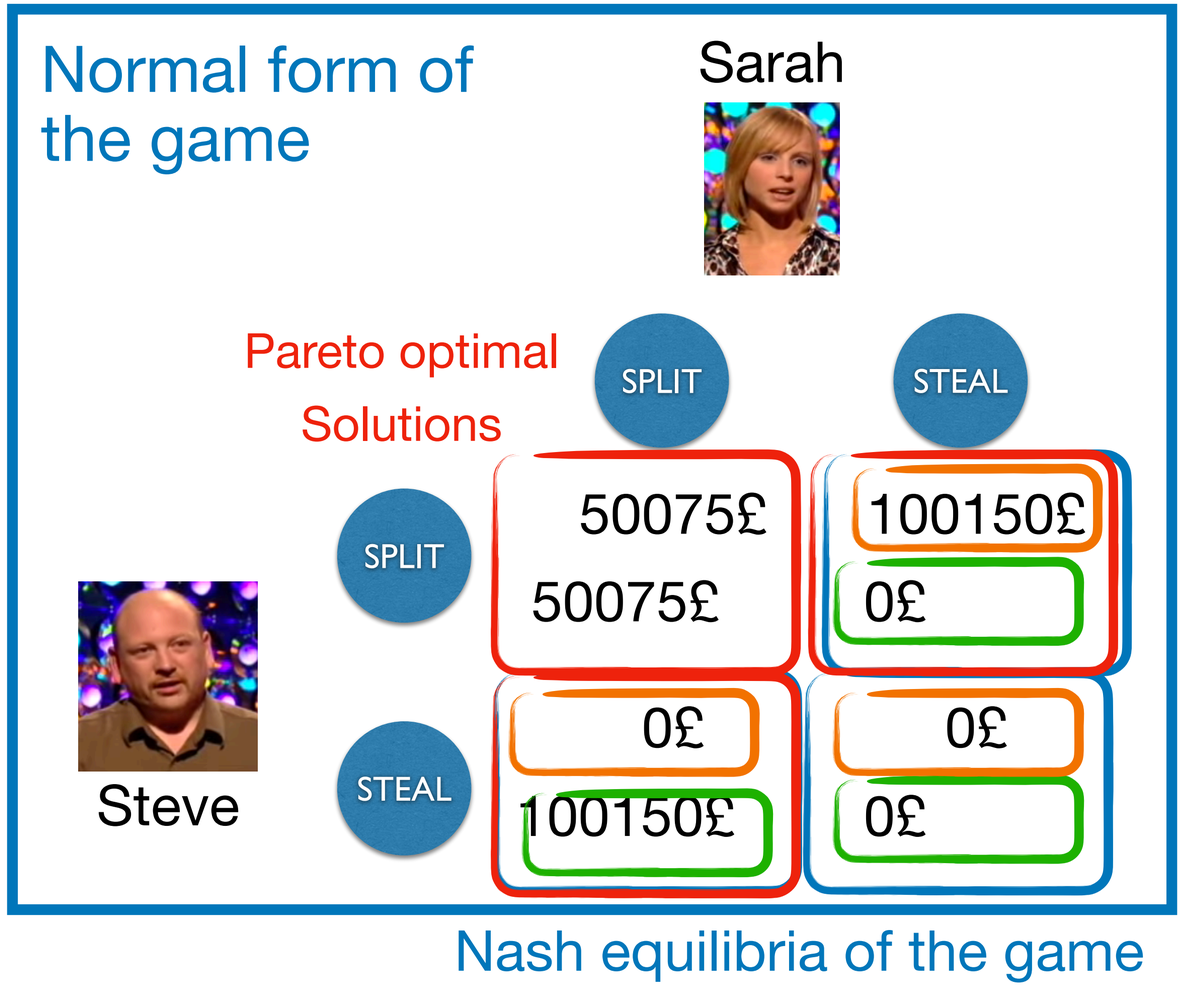
a_i^* is in $B_i(a_{-i}^*)$ for every player *i*

A best response is defined as:

$$B_i(a_{-i})=\{a_i \in A_i : u_i(a_i,a_{-i}) \geq u_i(a_i',a_{-i}) \ \forall \ a_i' \in A_i\}$$

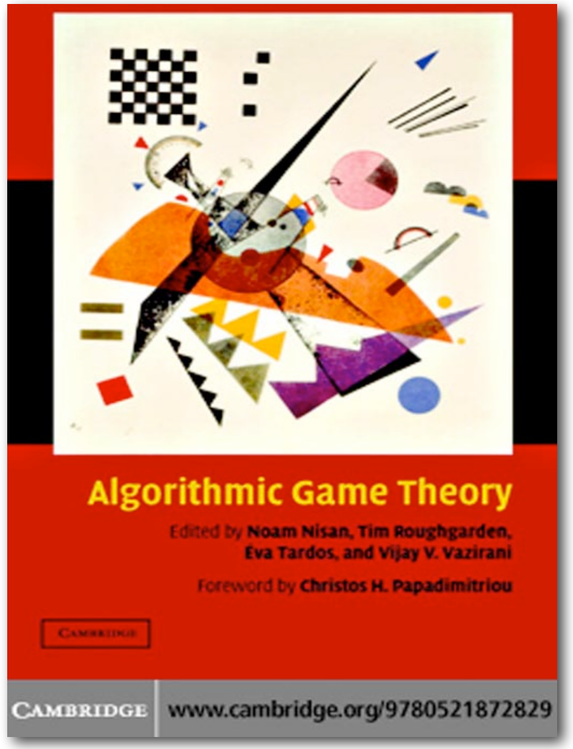
A Pareto optimal solution:

refers to an strategy profile in which it is impossible to improve the payoff of one player without worsening the payoff of another player



Sarah and Steve playing the golden balls game for 100150 pound

How to find all Nash equilibria for pairwise games with limited number of actions ?



Support finding

Vertex enumeration

Knowing the equilibria allows you to determine which one is preferred

Normal form of the game

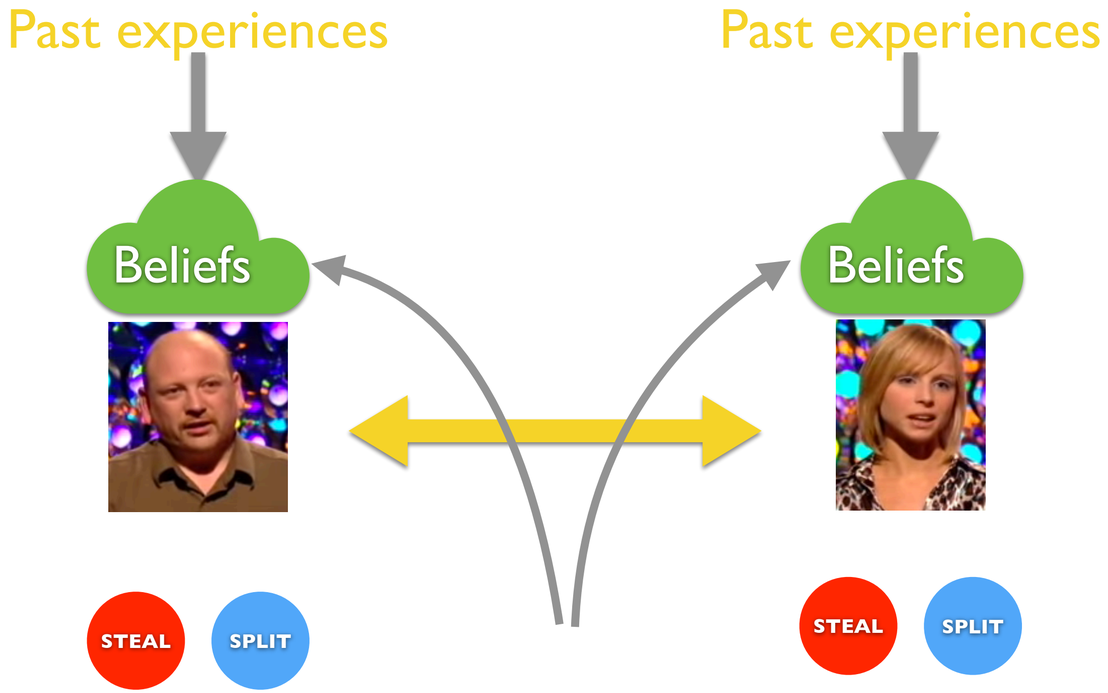

Steve

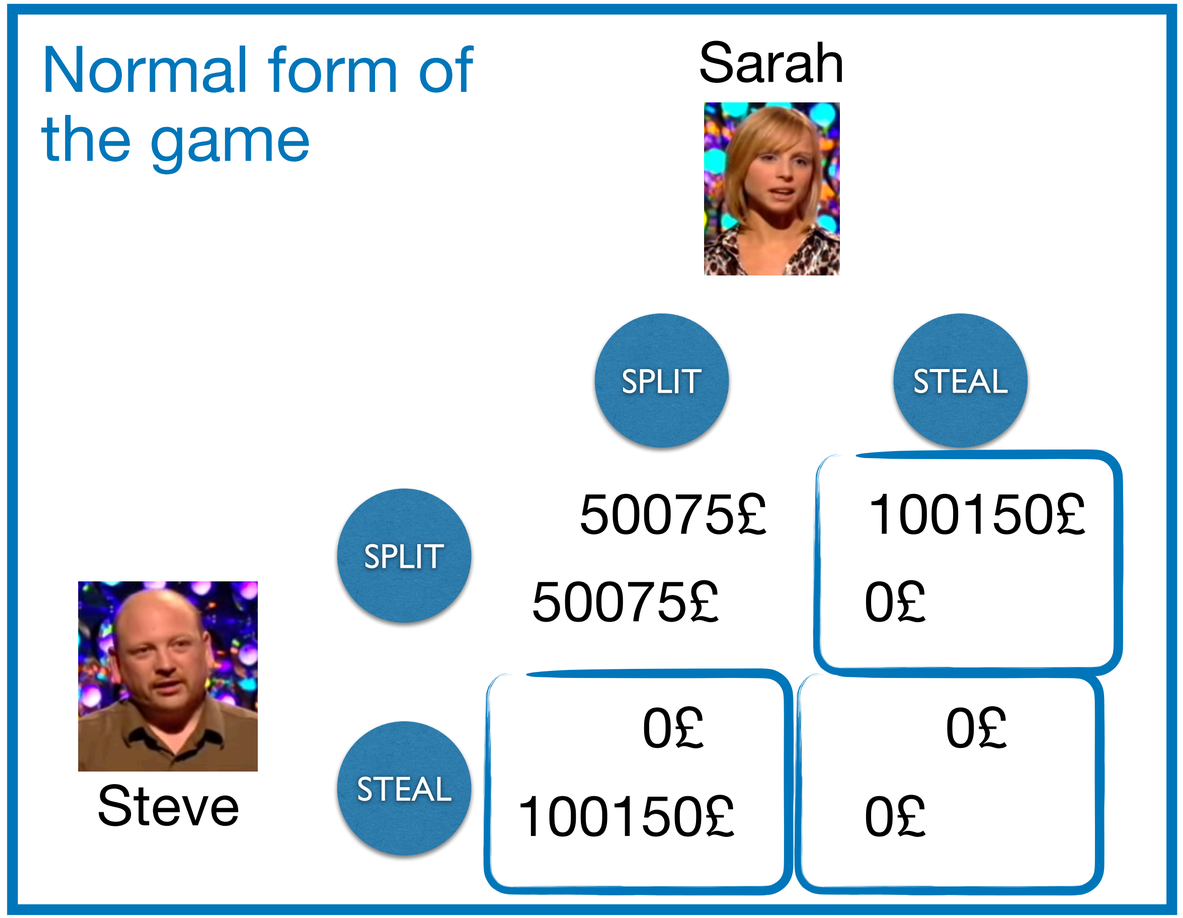

Sarah

Steve \ Sarah	SPLIT	STEAL
SPLIT	50075£ 50075£	100150£ 0£
STEAL	0£ 100150£	0£ 0£

Nash equilibria of the game

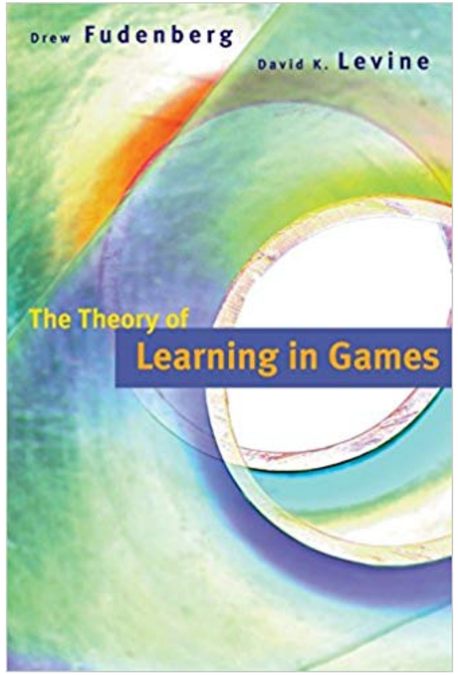
Sarah and Steve playing the golden balls game for 100150 pound





Nash equilibria of the game

Sarah and Steve playing the golden balls game for 100150 pound



Learning to reach an equilibrium

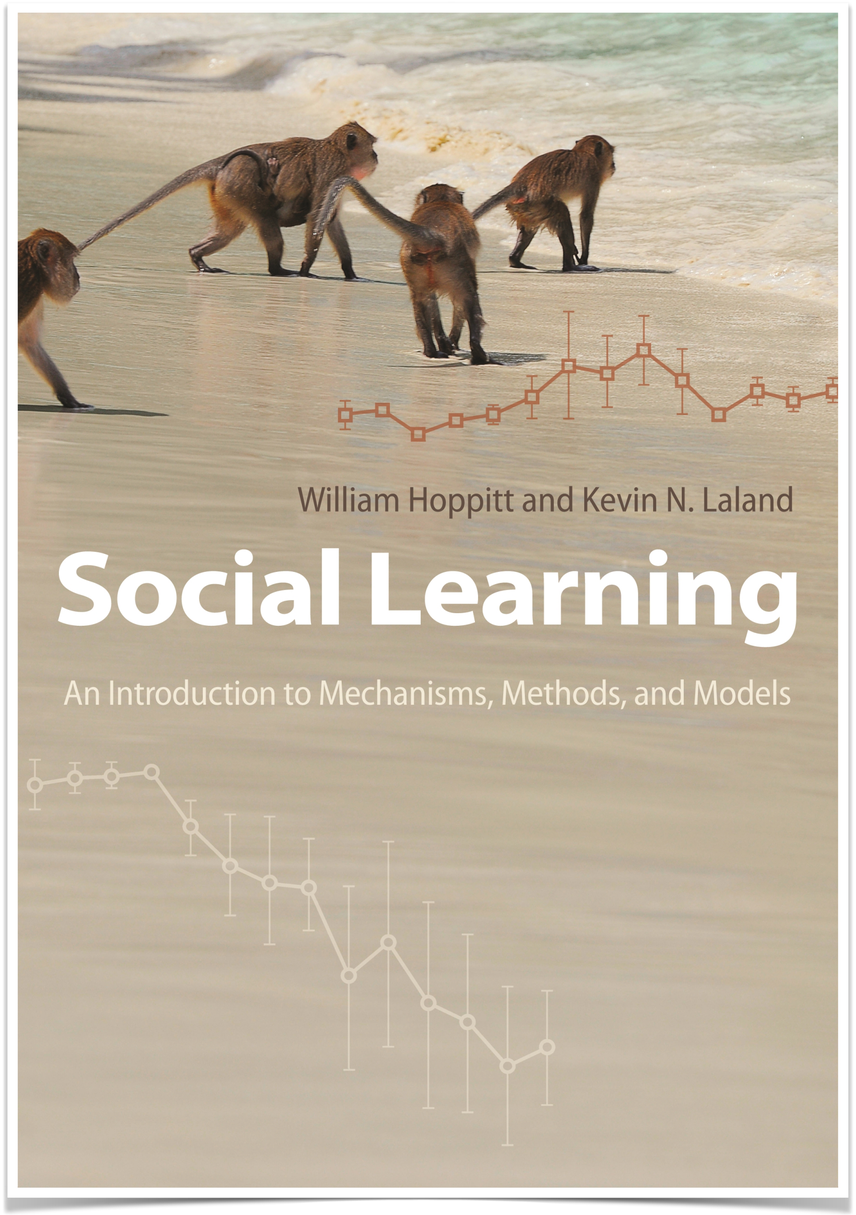
- Best response
- Fictitious play
- Roth-Erev learning, Experience-weight attraction learning, reinforcement learning
- Social/Evolutionary learning

Normal form of the game

Sarah

Steve \ Sarah	SPLIT	STEAL
SPLIT	50075£ 50075£	100150£ 0£
STEAL	0£ 100150£	0£ 0£

Steve



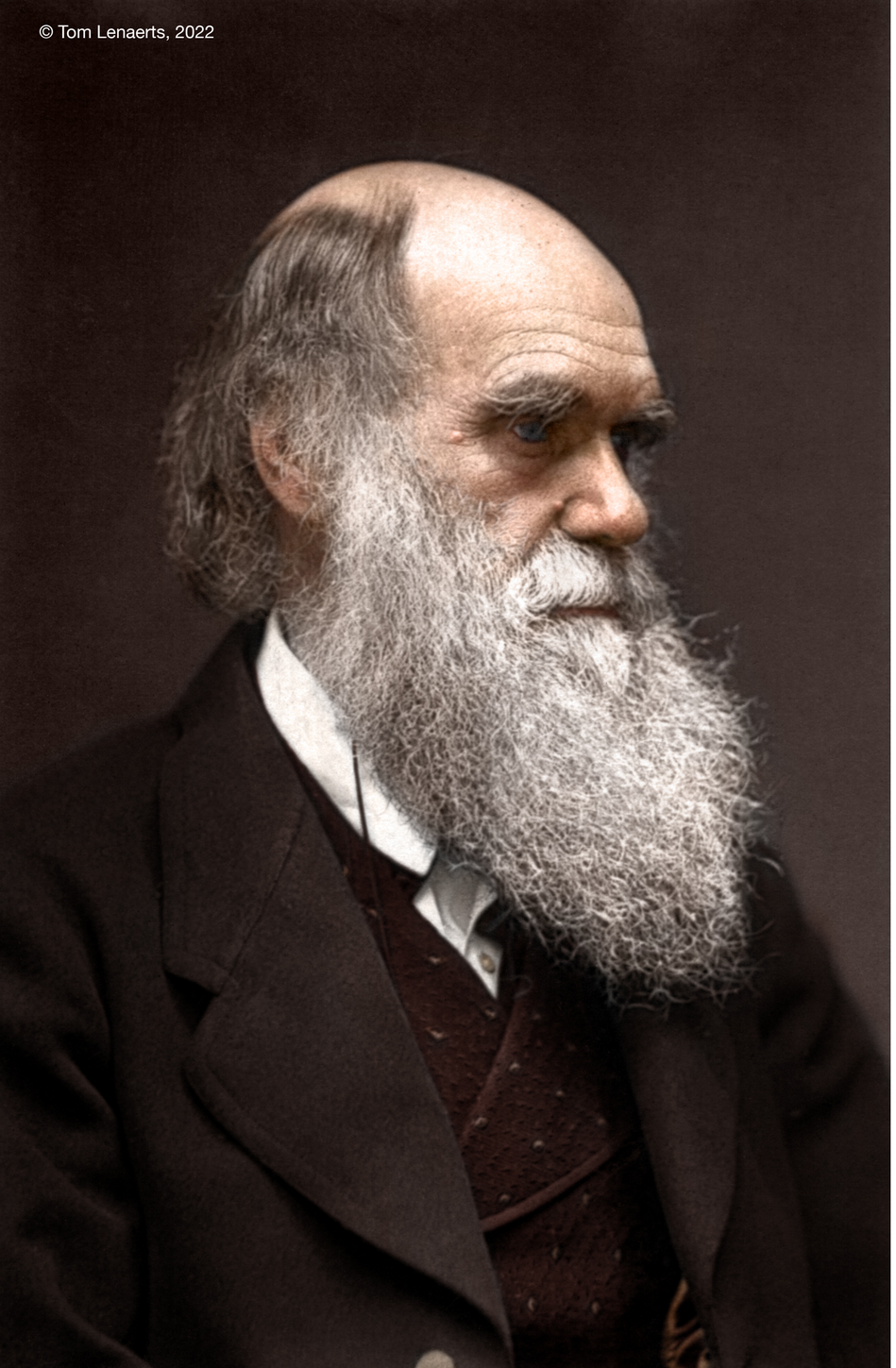
New behaviour is **acquired by observation/imitation**

Social learning is learning that is facilitated by observation , or interaction with, another individual or its products

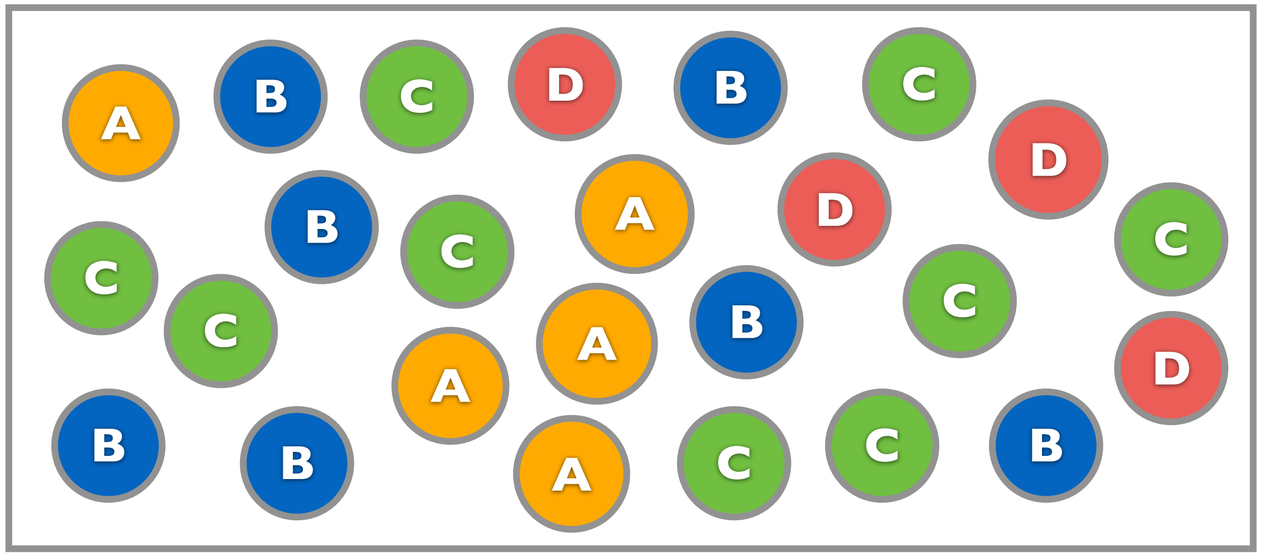
Evolutionary approach to model social learning

Reinforcement learning is also a form of observational learning, *with focus on the individual and happening at a different time-scale*

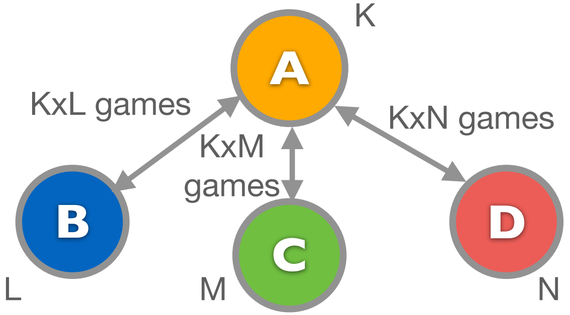
Norms and institutions occur at the population level, over generations



Non-rational players: **each player starts with one action**



Success **depends on the frequencies** of the different types of players



Darwinian competition driven by game success between players within populations

Social dilemmas

Golden balls game

Sarah

Steve \ Sarah	SPLIT	STEAL
SPLIT	50075£ / 50075£	100150£ / 0£
STEAL	0£ / 100150£	0£ / 0£

Cooperation

Defection

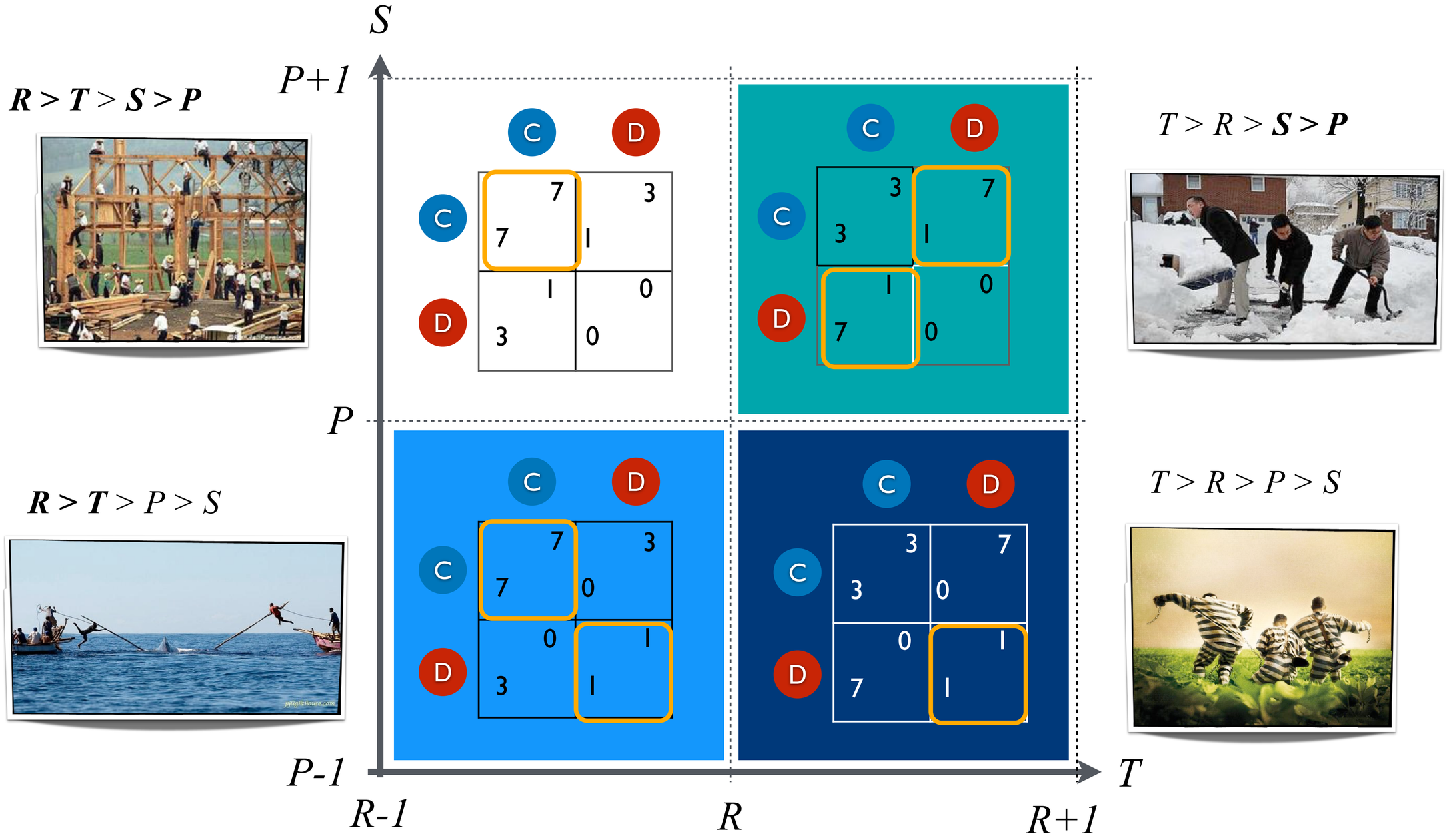
Prisoners Dilemma, T>R, P>S

Stag hunt, R>T

Snow drift, S>P

	C	D
C	R / R	S / T
D	T / S	P / P

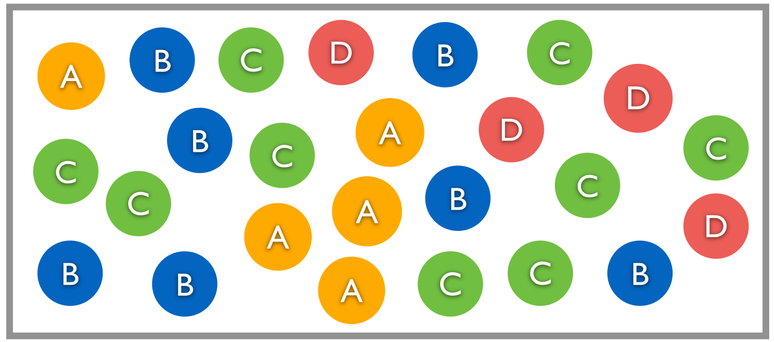
C.H. Coombs (1973) A reparameterization of the prisoner's dilemma game. Behavioral Science 18:424-428

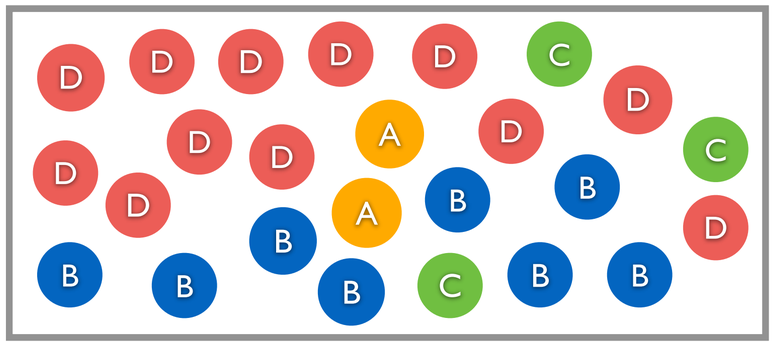


Darwinian competition

Imitate the best

Captures the interplay between the individual and the collective

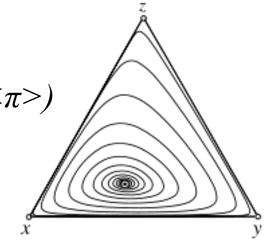




Evolutionary dynamics (Φ)

Replication dynamics

$\dot{x}_i = x_i(\pi_i - \langle\pi\rangle)$



(Nonlinear) dynamical systems

N→∞

(stochastic)
Birth-death (Moran) process

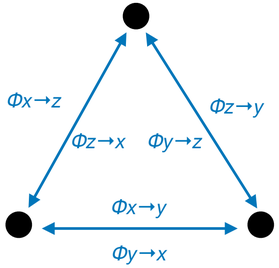
Act, Learn to act, Reason about act

(Agent-based simulation)

N finite

μ << 1

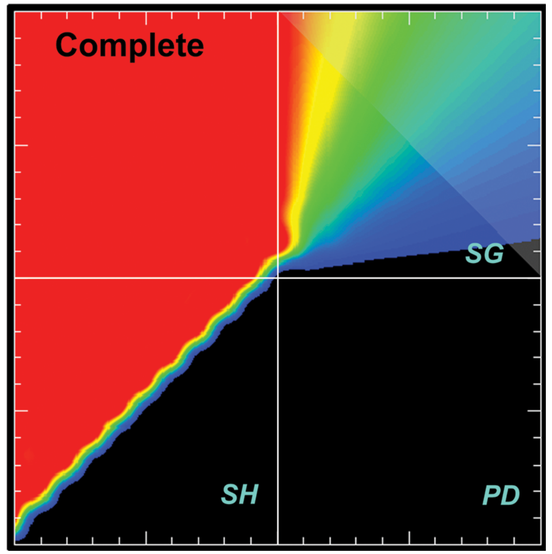
(stochastic)
Small mutation approximation



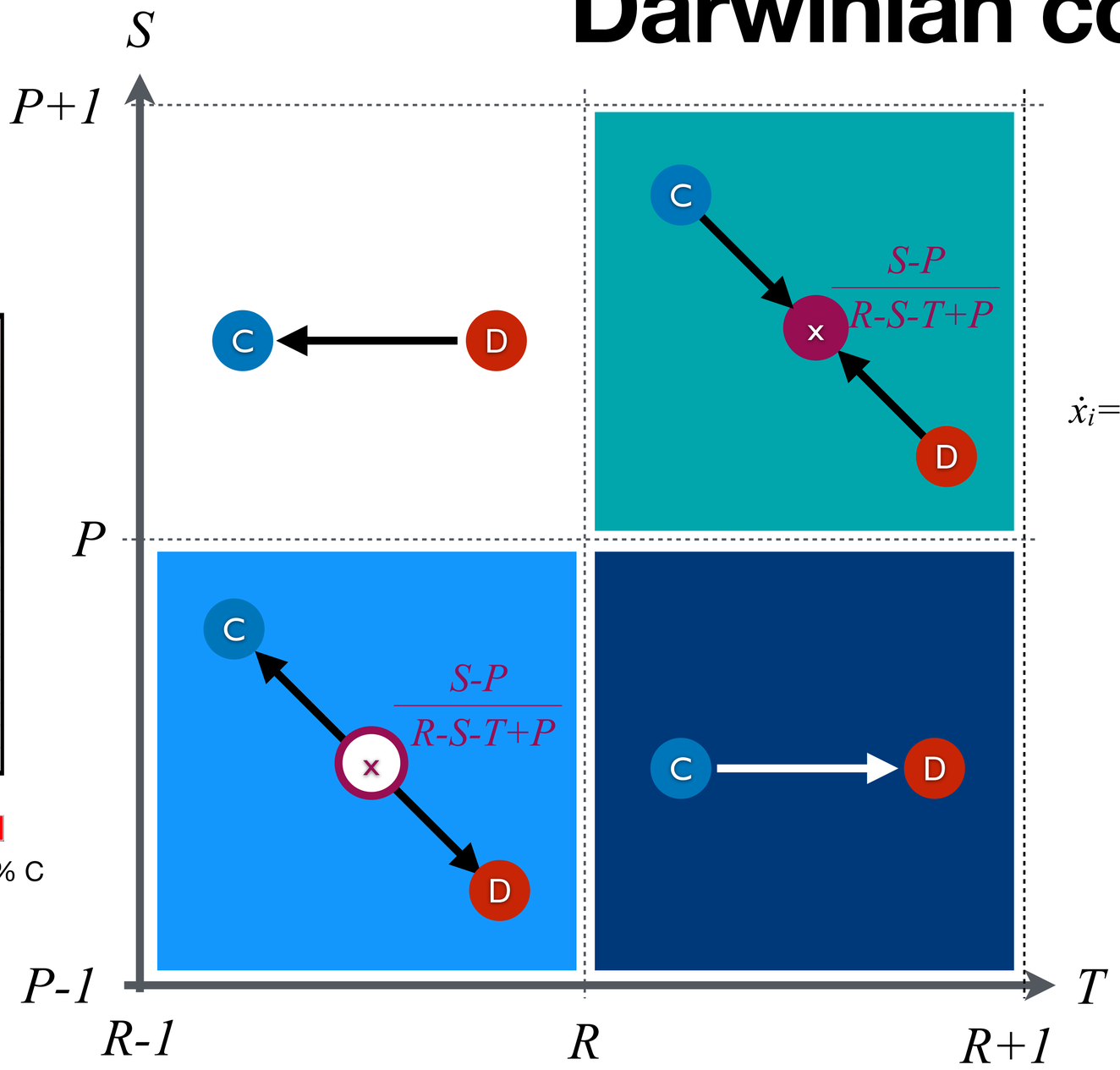
(Reduced Markov chains)

Adopted from Arne Traulsen

Darwinian competition

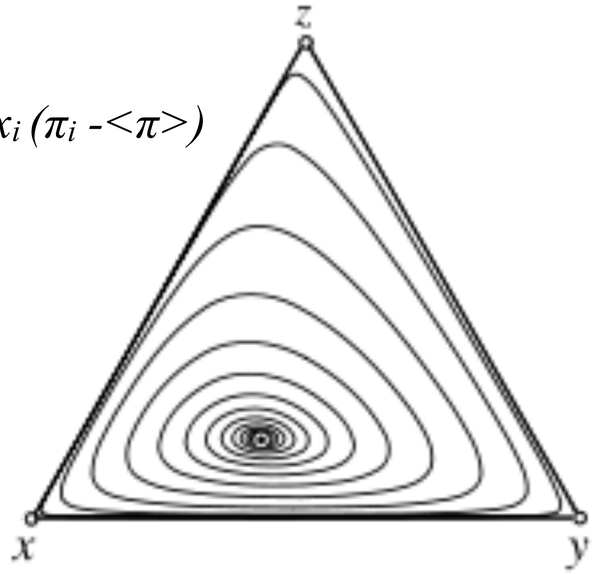


0% C 100% C



Replication dynamics

$\dot{x}_i = x_i(\pi_i - \langle\pi\rangle)$



Mechanisms of cooperation

Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 1331–1335, February 1982
Population Biology

Assortment of encounters and evolution of cooperativeness

(altruism/evolutionary stable strategies/assortative meetings)

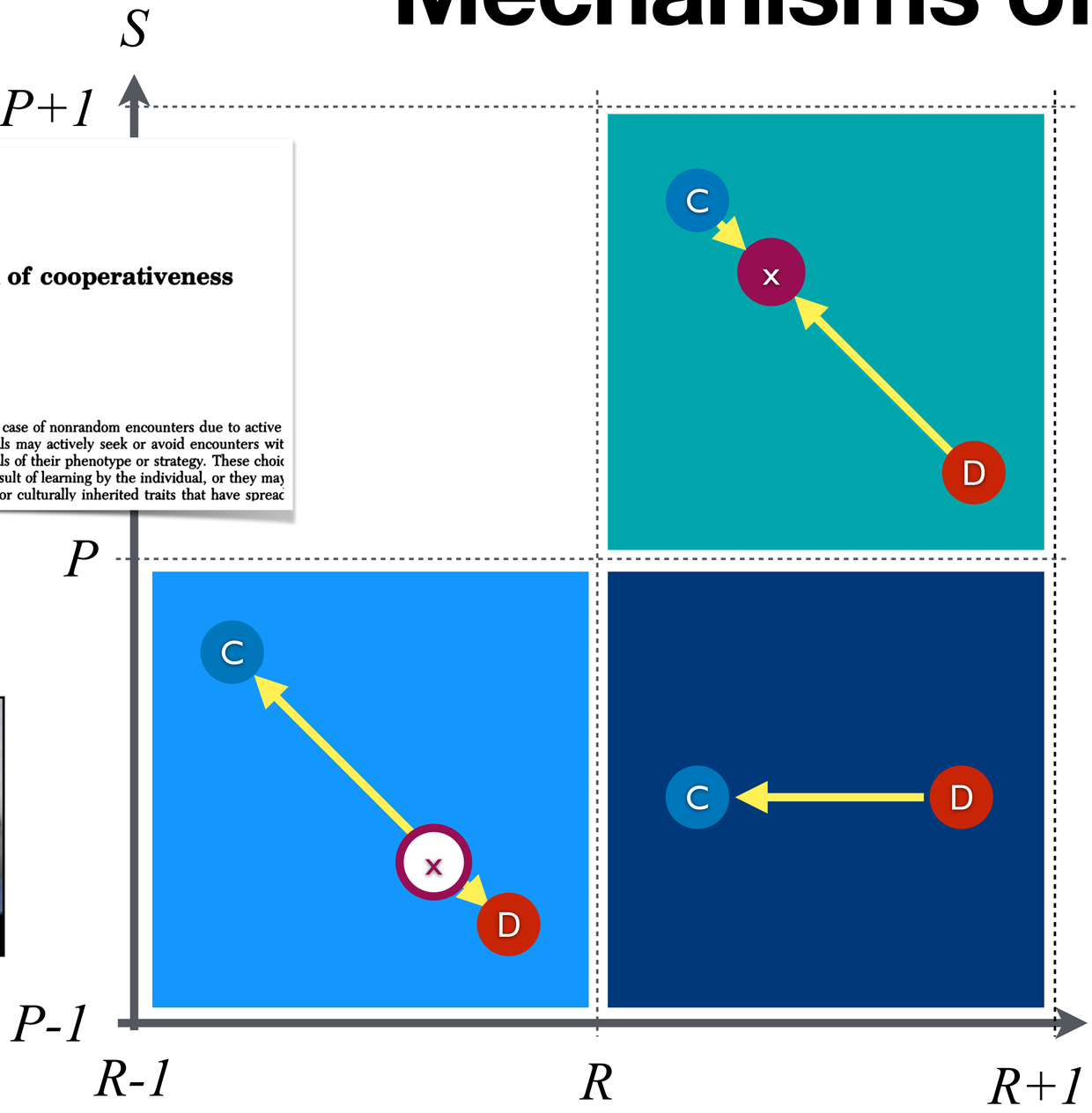
ILAN ESHEL† AND L. L. CAVALLI-SFORZA

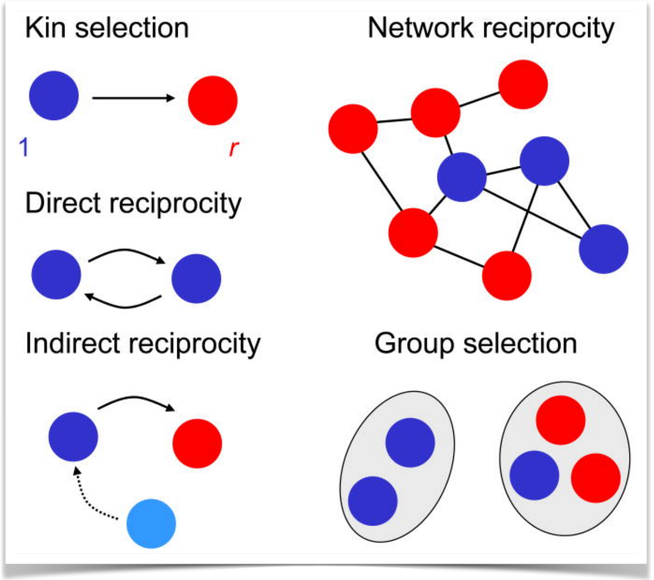
Departments of Mathematics and Genetics, Stanford University, Stanford, California 94305

Contributed by L. L. Cavalli-Sforza, October 13, 1981

ABSTRACT **The method of evolutionary stable strategies (ESS), in its current form, is confronted with a difficulty when it tries to explain how some social behaviors initiate their evolution. We show that this difficulty may be removed by changing the assumption made tacitly in game theory (and in ESS) of randomness**

In the case of nonrandom encounters due to active individuals may actively seek or avoid encounters wit individuals of their phenotype or strategy. These choic be the result of learning by the individual, or they may netically or culturally inherited traits that have spread





Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science*, *314*(5805), 1560-1563

Taylor, C., & Nowak, M. A. (2007). Transforming the dilemma. *Evolution: International Journal of Organic Evolution*, *61*(10), 2281-2292.

Understanding

AI needs a theory of mind, both affective and cognitive,

Brand New

Evolution of a Theory of Mind

Tom Lenaerts[a,b,c], Jorge M. Pacheco[d,e,f], and Francisco C. Santos[f,g]

[a] Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium

[b] Artificial Intelligence Lab, Vrije Universiteit Brussel, 1050, Brussels, Belgium

[c] Center for Human-Compatible AI, University of California, Berkeley, 94702 Berkeley, USA.

Communication

Credibly and explicitly share information,

CEPR Discussion Paper 17481 CEPR PRESS

Does voluntary information disclosure lead to less cooperation than mandatory disclosure? Evidence from a sequential prisoner's dilemma experiment.

Georg Kirchsteiger, Tom Lenaerts, Remi Suchon

July 19, 2022
Industrial Organization

Commitment

Have the capacity to uphold promises and

SCIENTIFIC REPORTS

OPEN

Good Agreements Make Good Friends

The Anh Han[1,2], Luís Moniz Pereira[3], Francisco C. Santos[4,5] & Tom Lenaerts[1,2]

[1]AI lab, Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium, [2]MLG, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium, [3]Centro de Inteligência Artificial (CENTRIA), Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal, [4]INESC-ID and Instituto Superior Ténico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal, [5]ATP-group, CMAF, Instituto para a Investigação Interdisciplinar, P-1649-003 Lisboa Codex, Portugal.

SUBJECT AREAS:
BIOLOGICAL PHYSICS
BEHAVIOURAL METHODS
EVOLUTIONARY THEORY
SOCIAL EVOLUTION

Norms and institutions

Needs social supervision so that shared beliefs and rules are followed

www.nature.com/scientificreports

SCIENTIFIC REPORTS

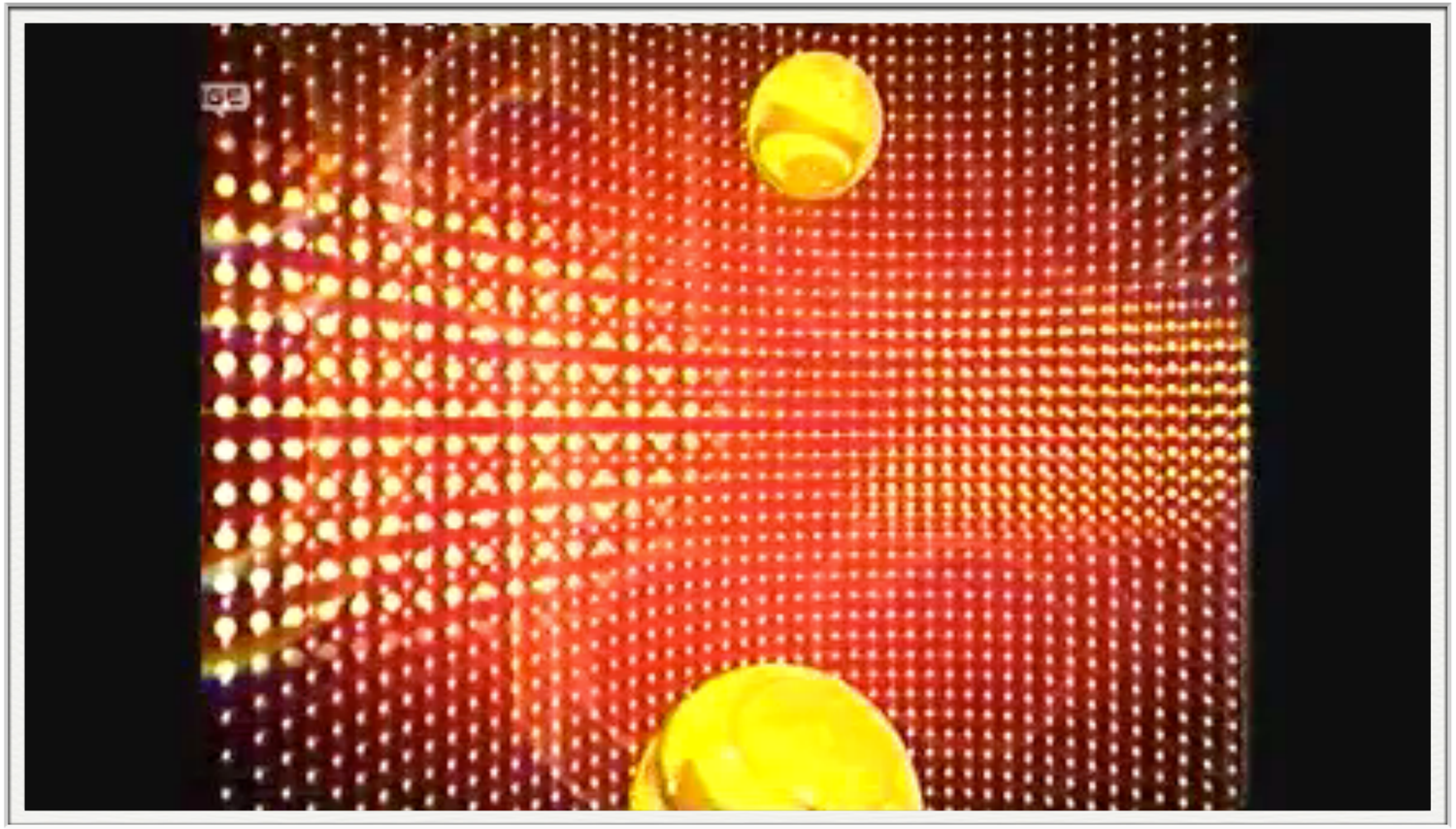
OPEN

Apology and forgiveness evolve to resolve failures in cooperative agreements

Received: 23 February 2015
Accepted: 22 April 2015
Published: xx xx xxxx

Luis A. Martinez-Vaquero[1,2], The Anh Han[3], Luís Moniz Pereira[4] & Tom Lenaerts[1,2]

Making agreements on how to behave has been shown to be an evolutionarily viable strategy in one-shot social dilemmas. However, in many situations agreements aim to establish long-term



“Golden Balls is a British daytime game show which was presented by Jasper Carrott. It was broadcast on the ITV network from 18 June 2007 to 18 December 2009. It was filmed at the BBC Television Centre. Golden Balls Ltd licensed their name to Endemol for the game show and merchandise.” [Wikipedia Oct. 2020]

Nick

Share (or not) 13600£

Abraham

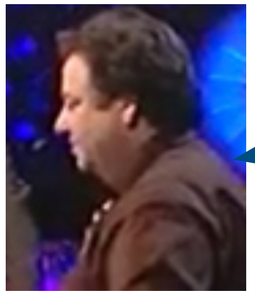
SPLIT

STEAL

Abraham \ Nick	SPLIT	STEAL
SPLIT	6800£, 6800£	0£, 13600£
STEAL	13600£, 0£	0£, 0£

Nash equilibria

Nick



I will steal, so if you want to see any money, you need to choose split and I will share with you after the show

Share (or not) 13600£

If I want to see any money I need to split

Abraham

SPLIT

STEAL

SPLIT

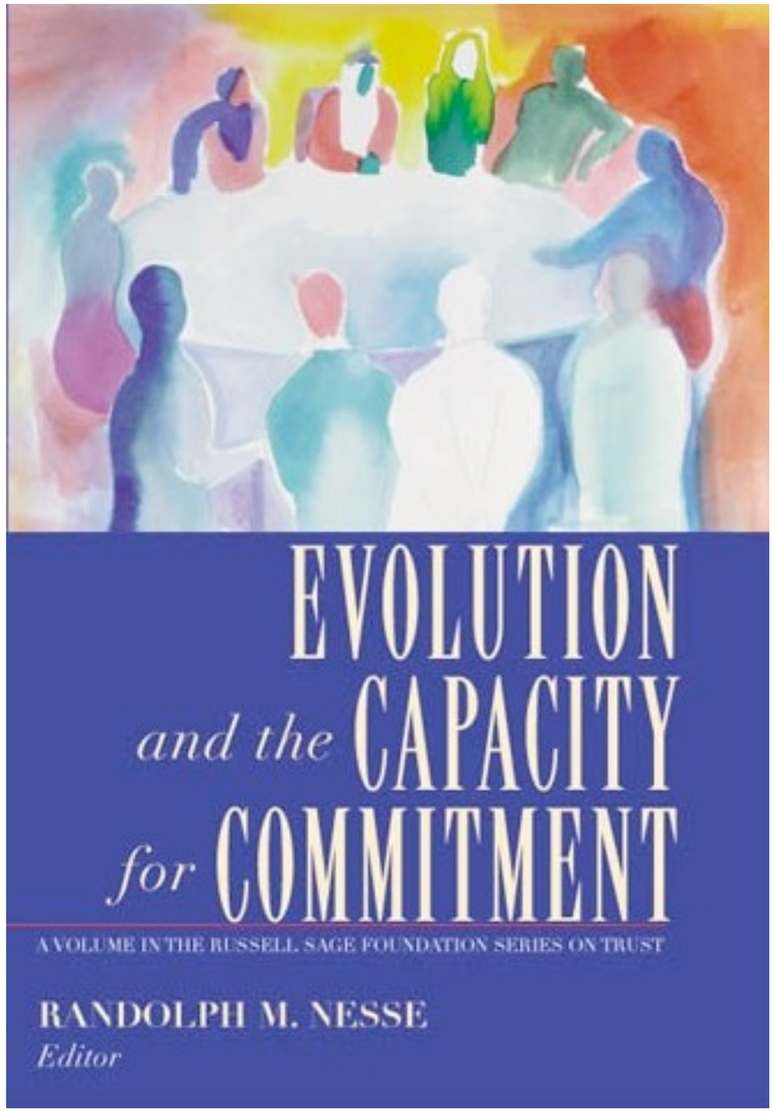
STEAL

	6800£	13600£
6800£		0£
0£		0£
13600£		0£

Nick **commits** unilaterally to steal

By **committing**, he removes choices from Abraham

Needs to make the sharing after the show credible!



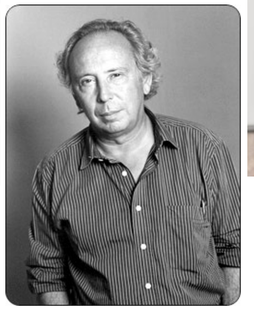
"A **commitment** is an act or signal that gives up options in order to **influence someone's behaviour** by changing incentives and expectations"

"Commitments can be **promises** to **help**, or **threats** to **harm**"

"They can be **enforced** by external incentives, but also by some combination of **reputation** and **emotion**"

"Our (cognitive) capacity for commitment may have **evolved by natural selection**"

Our results

- Costly commitments with compensations work well to generate cooperation

  Evolution selects for this behaviour

- More effective than costly punishment (see paper)

- **Apologising** and **forgiving** appear to be key for stable prosocial relations

  As long as the apology is sincere enough (cost)

- ***Cooperative AI*** *; Can we use autonomous agents as a way to commit to certain behaviour ?*

  Delegation of decision-making from human to agent

Can autonomous agents act as a **commitment devices**?

Collective risk dilemma

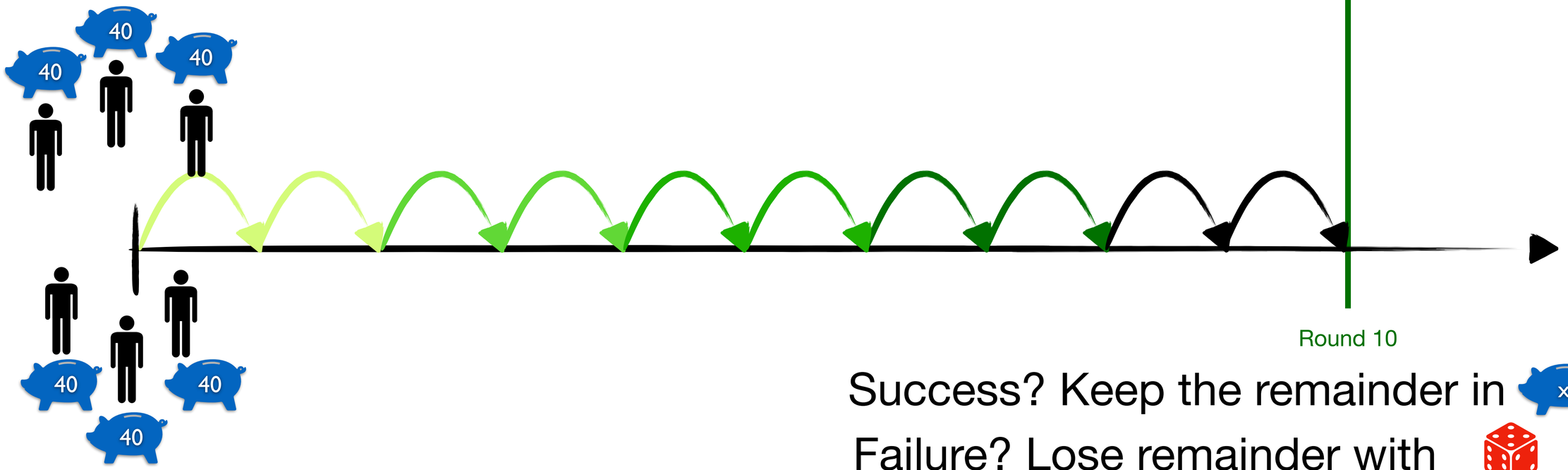
6 **players** :

Each player receives 40

Actions: give in each round {0,2,4}

Repeat 10 times

Goal = 120



Success? Keep the remainder in x

Failure? Lose remainder with

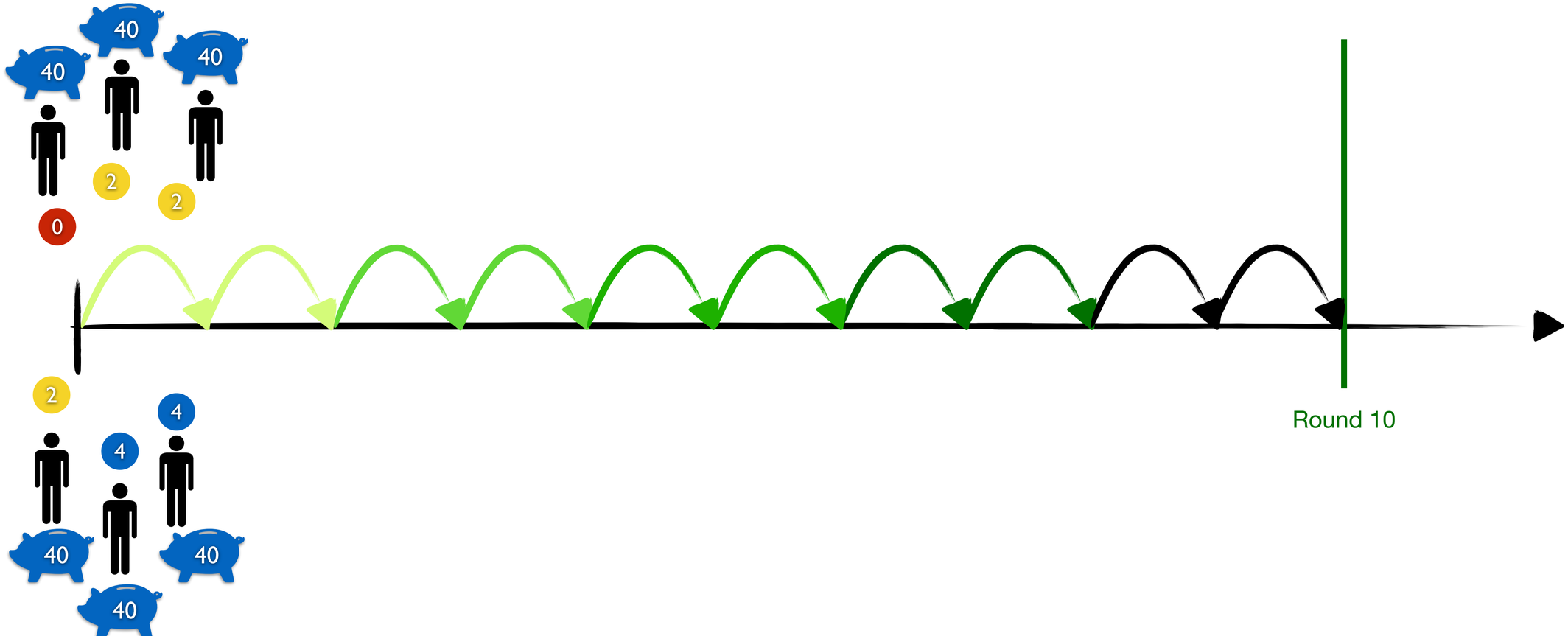
M. Milinski et. al, “The collective-risk social dilemma and the prevention of simulated dangerous climate change.,” PNAS (2008).

Collective risk dilemma

6 **players**

Actions: give in each round {0,2,4}

Repeat 10 times

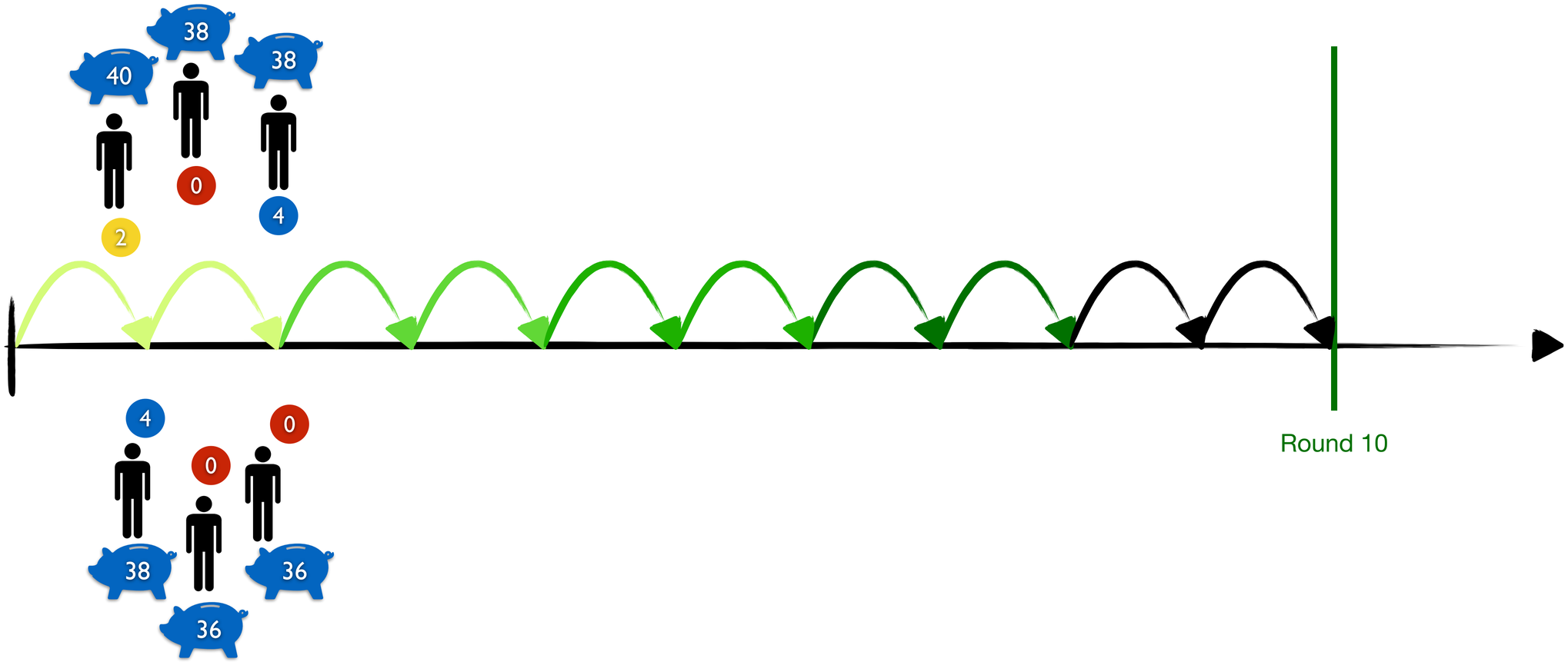


Collective risk dilemma

6 **players** 

Actions: give in each round {0,2,4}

Repeat 10 times

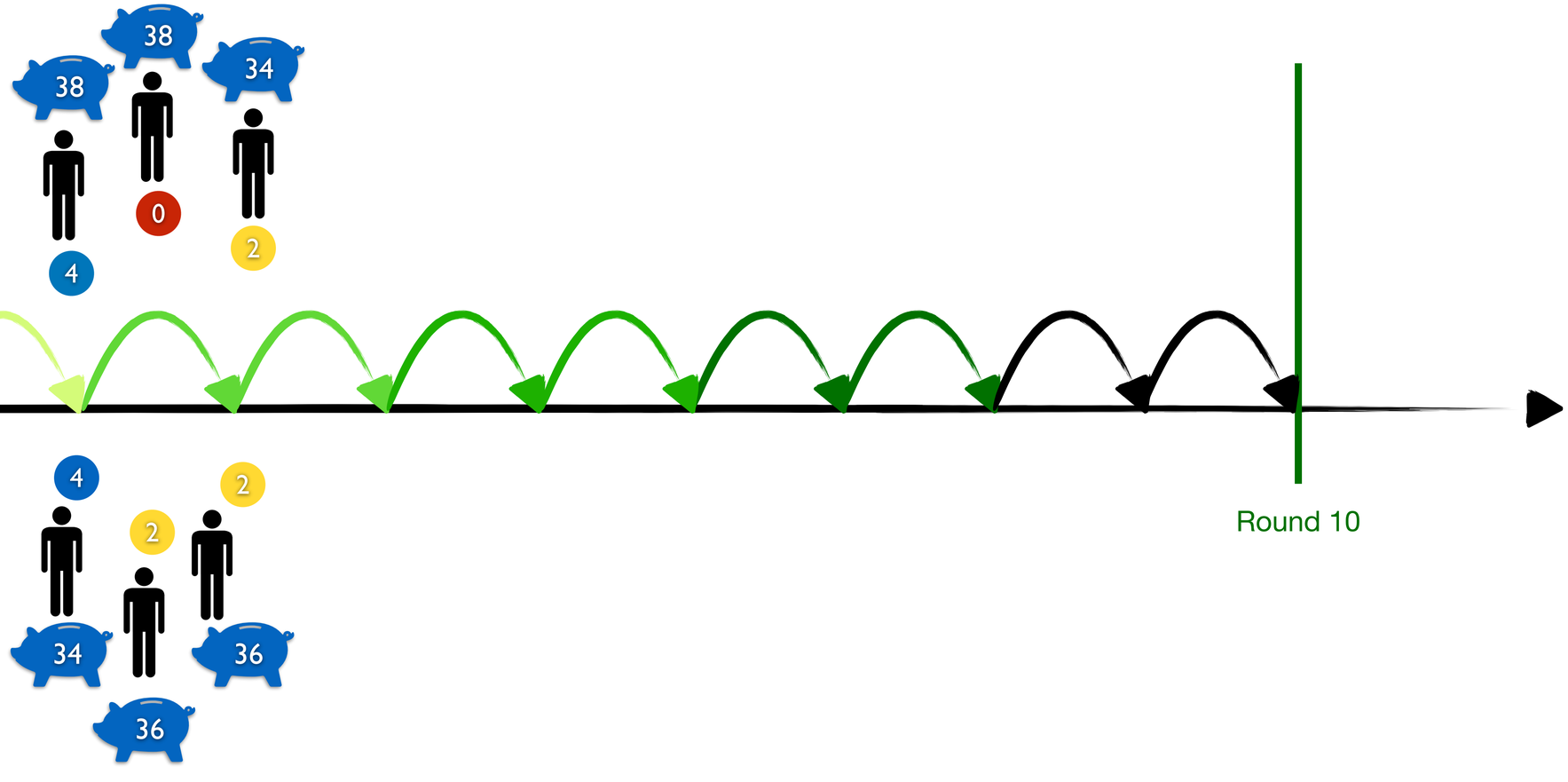


Collective risk dilemma

6 **players**

Actions: give in each round {0,2,4}

Repeat 10 times

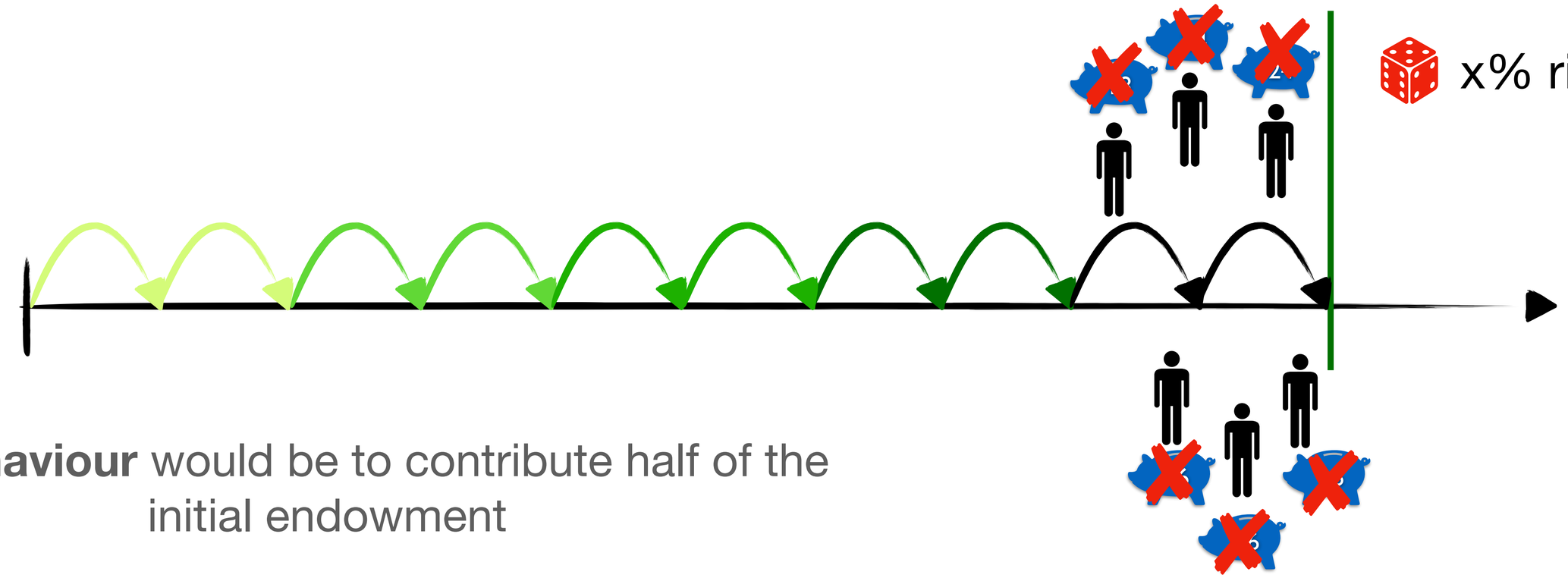


Collective risk dilemma

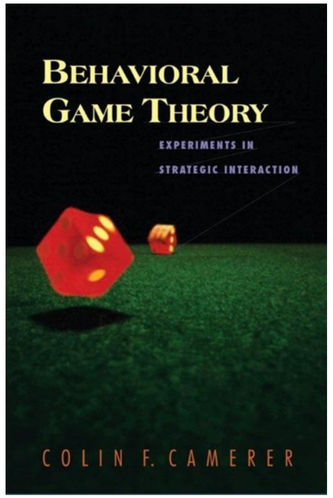
6 **players**

Actions: give in each round {0,2,4}

Repeat 10 times



Fair behaviour would be to contribute half of the initial endowment



round 2 of 10

Donations of the previous round

You	Other members of the group			
2	0	2	2	0	2

How many EMUs do you want to contribute to the public account?

Select one of the following options.

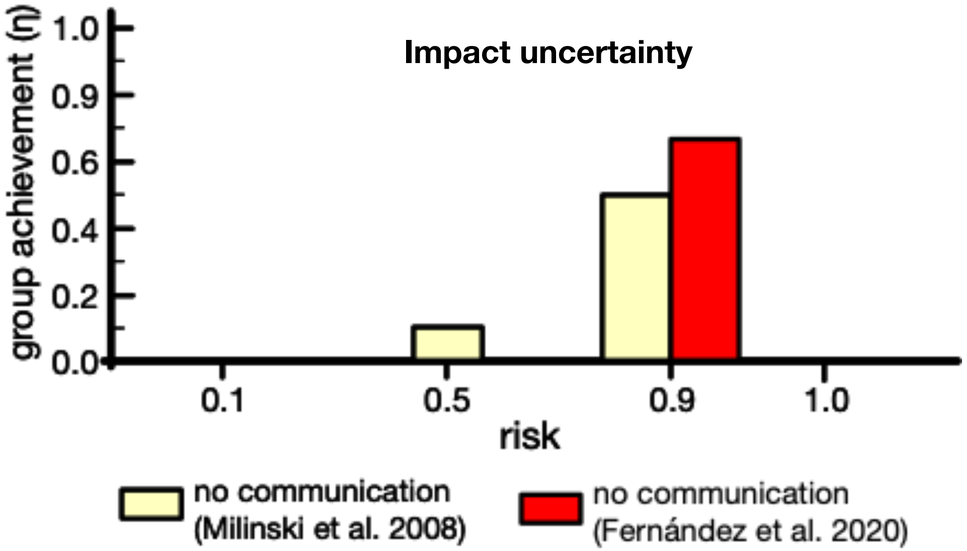
0 2 4

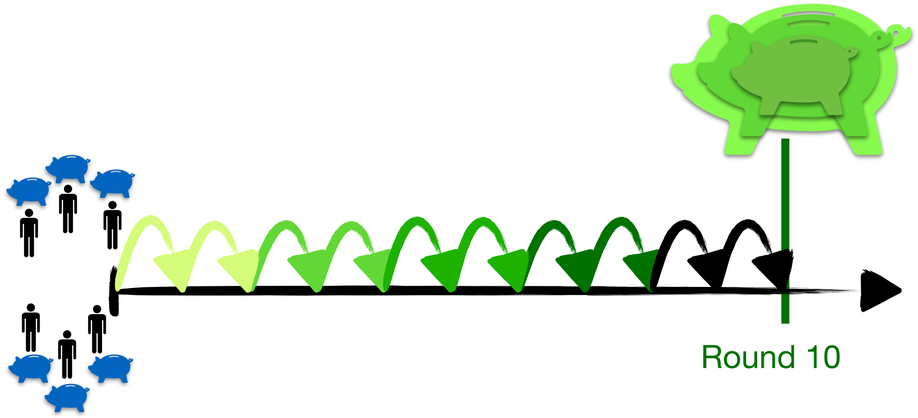
Time left

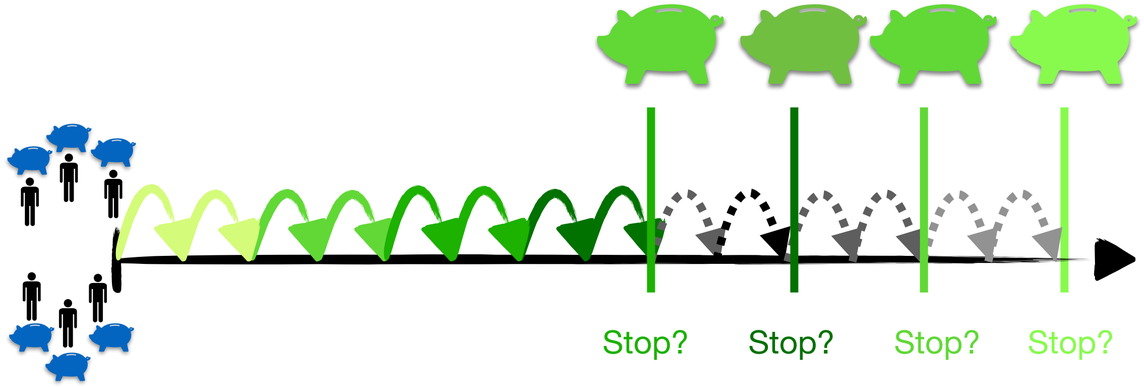
00:53

Personal Account

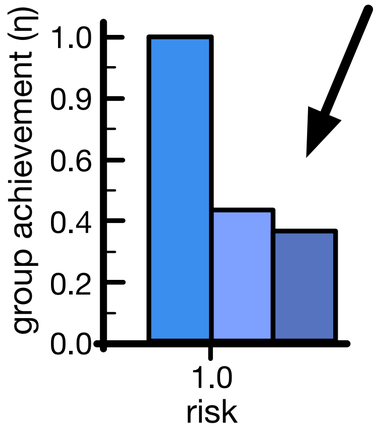
38 EMUs







Threshold uncertainty



communication & threshold uncertainty with uniform distribution (Dannenberg et al. 2014)

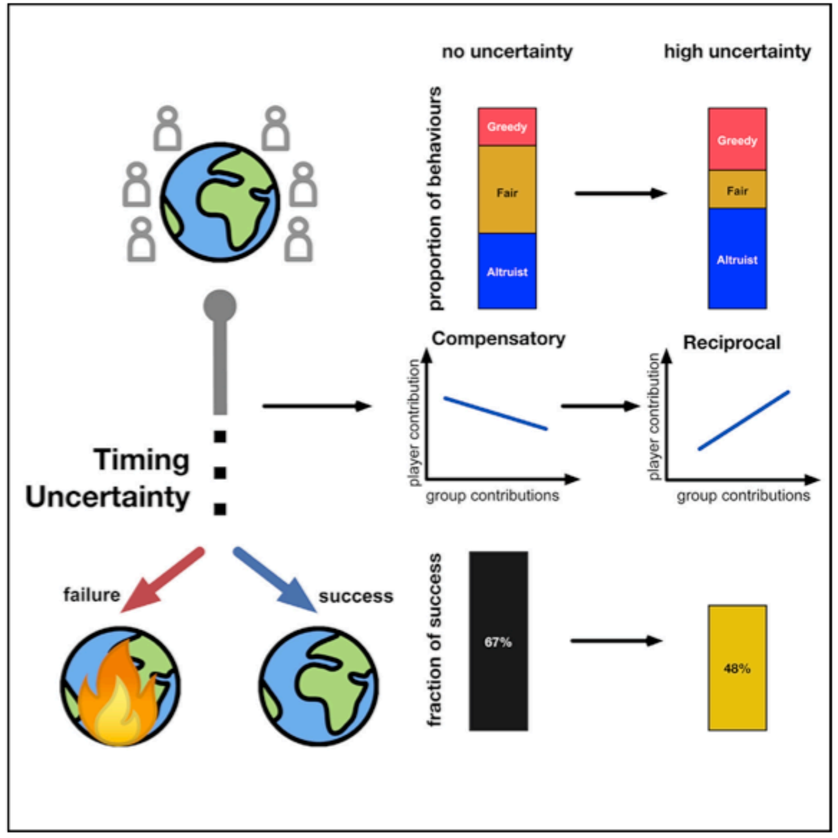
communication & threshold uncertainty with unknown distribution (Dannenberg et al. 2014)

iScience

CellPress
OPEN ACCESS

Article

Timing Uncertainty in Collective Risk Dilemmas Encourages Group Reciprocation and Polarization



Elias Fernández Domingos, Jelena Grujić, Juan C. Burguillo, Georg Kirchsteiger, Francisco C. Santos, Tom Lenaerts

franciscocsantos@tecnico.ulisboa.pt (F.C.S.)
tlenaert@ulb.ac.be (T.L.)

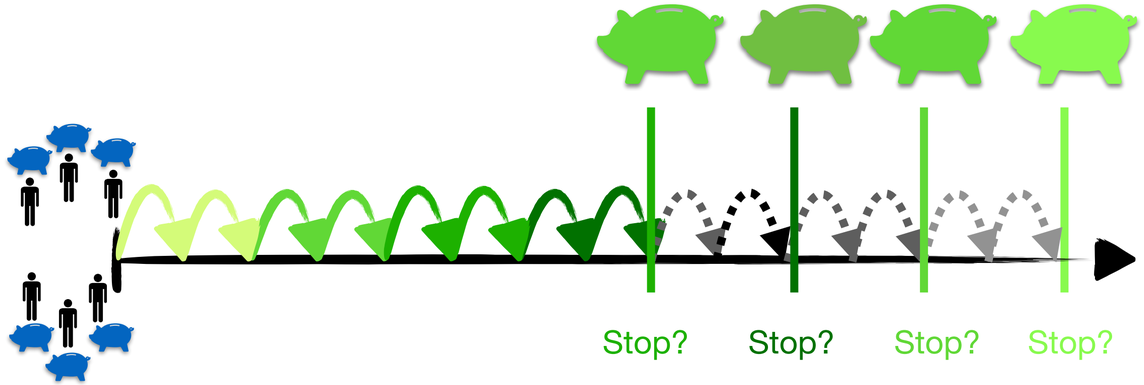
HIGHLIGHTS

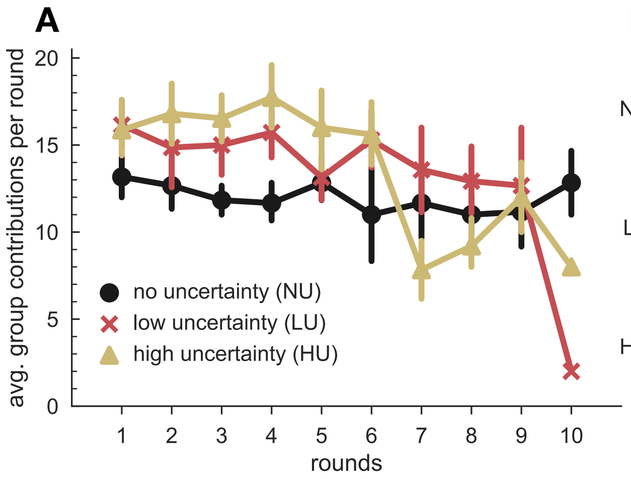
Timing uncertainty influences experimental observations in the collective risk game

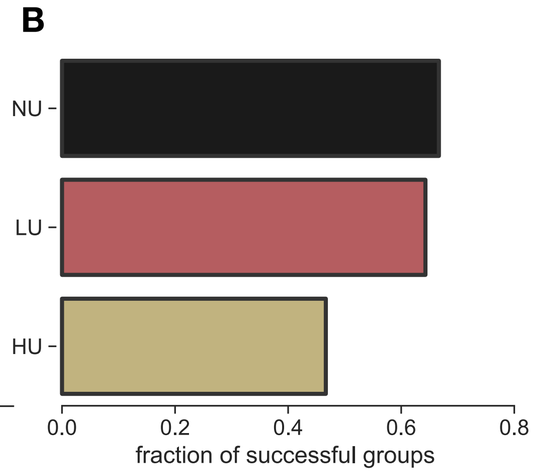
It induces subjects to contribute earlier and in a polarized manner

Successful players adopt reciprocal strategies, responding in kind to past actions

Coordination gets more difficult under high timing uncertainty







Humans appear to have problems coordinating their actions, even when the risk is high

Additional sources of uncertainty are detrimental to success

In timing uncertainty

- actions become polarised
- strategies switch from compensation to reciprocal

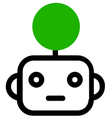
Helpful versus fairness-seeking

Both **EGT** and **population-based RL** models align with the observations

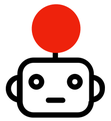
Will AI delegates solve the problem?

Delegation

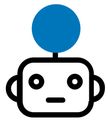
Select an agent that will play the game for you



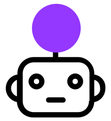
Always give 4



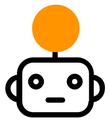
Always give 0



Always give 2



Give 4 when the group gave less than 10 in the previous round, otherwise 0



Give 0 when the group gave less than 10 in the previous round, otherwise 4

Customize

Program your preferred behaviour in this template agent

For i: 1- 10

If (i==1) give $\mathbf{a_0}$

Else

If (prev > **T**) give $\mathbf{a_a}$

Else If (prev < **T**) give $\mathbf{a_b}$

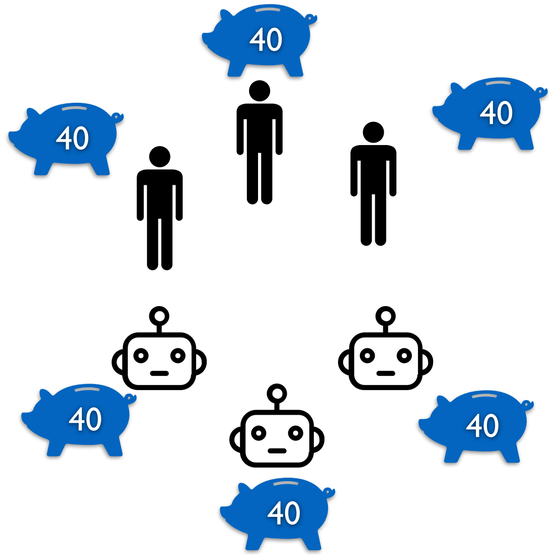
Else give $\mathbf{a_m}$

Each human participant defines the values for the parameters:

T, a_0, a_a, a_b and a_m

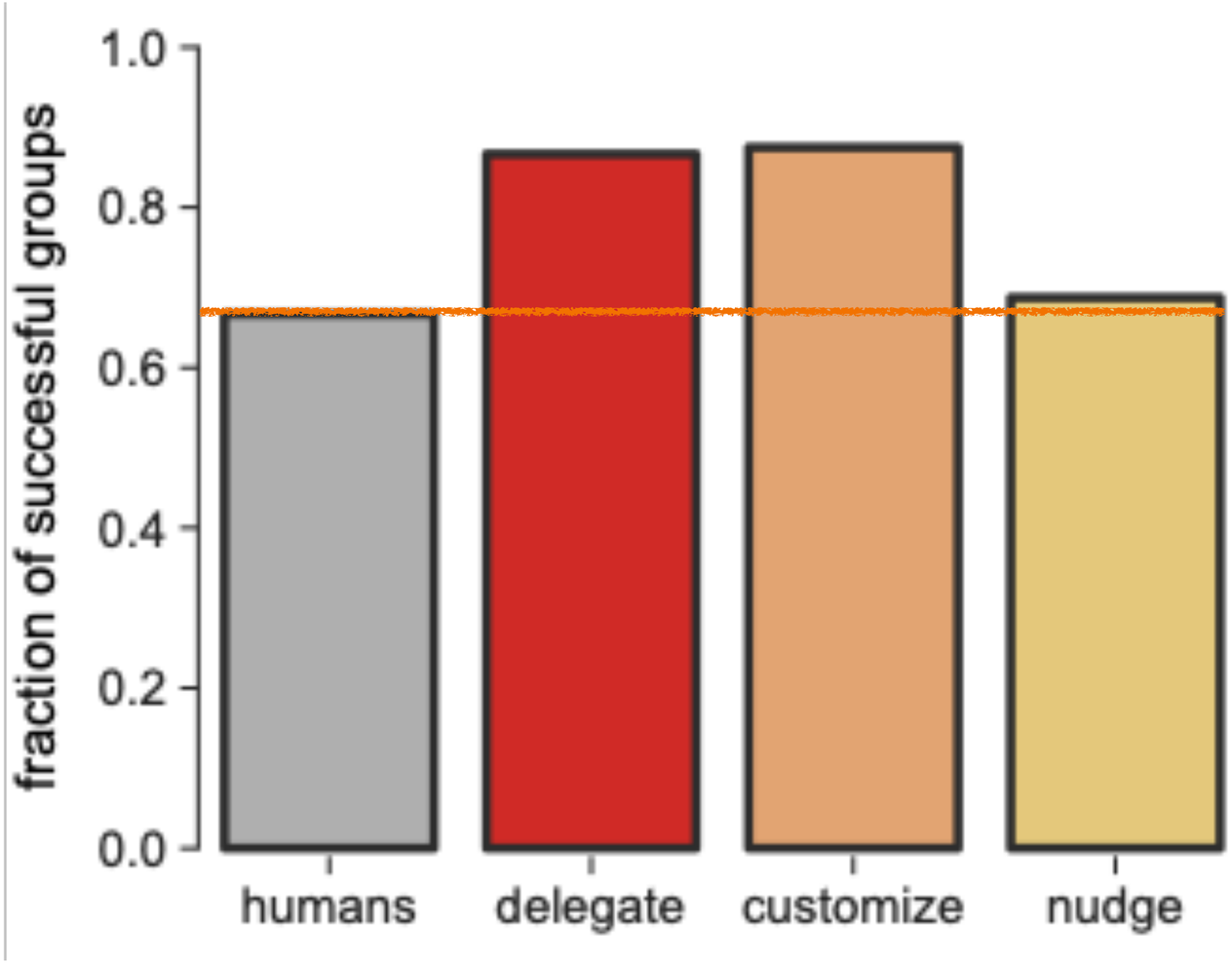
Nudge

Play game with half humans and half agents



Agents most successful in achieving the goal from the programming experiment

Success increases significantly when actions are delegated to agents



scientific reports

OPEN

Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma

Elias Fernández Domingos[1,2,8]✉, Inês Terrucha[2,3], Rémi Suchon[1,4], Jelena Grujić[1,2], Juan C. Burguillo[5], Francisco C. Santos[6]✉ & Tom Lenaerts[1,2,7,8]✉

Home assistant chat-bots, self-driving cars, drones or automated negotiation systems are some of the several examples of autonomous (artificial) agents that have pervaded our society. These agents enable the automation of multiple tasks, saving time and (human) effort. However, their presence in social settings raises the need for a better understanding of their effect on social interactions and how they may be used to enhance cooperation towards the public good, instead of hindering it. To this end, we present an experimental study of human delegation to autonomous agents and hybrid human-agent interactions centered on a non-linear public goods dilemma with uncertain returns in which participants face a collective risk. Our aim is to understand experimentally whether the presence of autonomous agents has a positive or negative impact on social behaviour, equality and cooperation in such a dilemma. Our results show that cooperation and group success increases when participants delegate their actions to an artificial agent that plays on their behalf. Yet, this positive effect is less pronounced when humans interact in hybrid human-agent groups, where we mostly observe that humans in successful hybrid groups make higher contributions earlier in the game. Also, we show that participants wrongly believe that artificial agents will contribute less to the collective effort. In general, our results suggest that delegation to autonomous agents has the potential to work as commitment devices, which prevent both the temptation to deviate to an alternate (less collectively good) course of action, as well as limiting responses based on betrayal aversion.

- Delegation in the CRD appears to increase success
- Delegation is trusted more when users can customise the agent
- When delegating to an AI/algorithm,
  - one commits to a certain course of actions
  - emotional responses to past behaviours do not play a role

  Removing fear of betrayal
- Agents are wrongfully considered to be less contributing
- **Are these conclusions generally true? Is there more to the story ?**

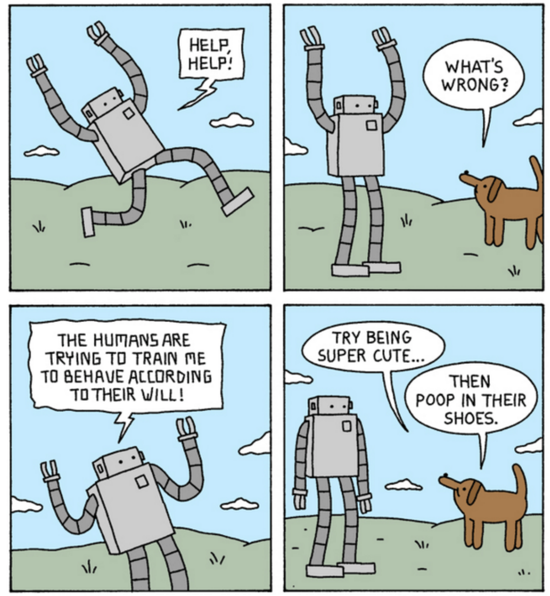
Take home message

- Cooperation is key concept which needs to be carefully studied, also in the context of AI ecosystems
- Don’t reinvent the wheel (or terminology), a lot of work has been done
- Work is needed on bringing the results of EGT closer to classic single-agent AI
- Experiments are needed to validate models but also to guide model design
- Simple benchmarks provide explainable solutions

Want to know more?

mlg.ulb.ac.be

Tom.Lenaerts@ulb.be



https://blog.rebellionresearch.com/blog/a-i-as-a-society-of-idiot-savants

tomlenaerts.be

MLG MACHINE LEARNING GROUP

ULB

ARTIFICIAL INTELLIGENCE RESEARCH GROUP

VUB

FARI Ai FOR THE COMMON GOOD INSTITUTE BRUSSELS

CHAI Center for Human-Compatible Artificial Intelligence

B² INTERUNIVERSITY INSTITUTE OF BIOINFORMATICS IN BRUSSELS