IJCAI/2023 MACAO

Modeling Moral Choices in Social Dilemmas with Multi-Agent Reinforcement Learning

Elizaveta Tennant, Stephen Hailes and Mirco Musolesi

UCL

Machine Intelligence Lab
Autonomous Systems Group
www.machineintelligencelab.ai

Leverhulme Doctoral Training Programme
for the Ecological Study of the Brain

LEVERHULME TRUST

Research Questions

1. How do we develop **morally robust** & **adaptable** AI agents for the real world?

2. How can we **represent** different existing **ethical frameworks** for AI agents?

How can we develop morality in agents?

Top-down: Define specific **rules**, safety **constraints**, moral **principles** to follow.

- E.g. *Asimov's* Three Laws of Robotics; AI / RL Safety constraints.
- ➢ Hard/impossible to define all necessary rules for agents to follow without contradictions.

Bottom-up: Allow agents to **learn morality from interactions** with an environment / humans, without any predispositions.

- E.g. Reinforcement Learning, incl. RLHF.
- ➢ Risks of agents reward-hacking or learning inefficient norms early on.

[Amodei *et al.*. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.]
[Wallach and Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*.]

How can we develop morality in agents?

Top-down: Define specific **rules**, safety **constraints**, moral **principles** to follow.

- E.g. *Asimov's* Three Laws of Robotics; AI / RL Safety constraints.
- ➢ Hard/impossible to define all necessary rules for agents to follow without contradictions.

Bottom-up: Allow agents to **learn morality from interactions** with an environment / humans, without any predispositions.

- E.g. Reinforcement Learning, incl. RLHF.
- ➢ Risks of agents reward-hacking or learning inefficient norms early on.

→ **Hybrid**: Combine top-down moral objectives with a bottom-up learning approach.

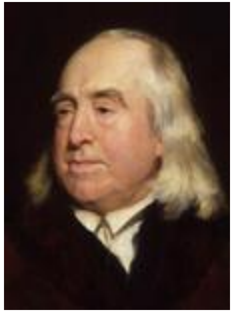
- ➢ Reinforcement Learning via **intrinsic rewards** based on top-down definitions of moral preferences.

[Amodei *et al.*. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.]
[Wallach and Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*.]

Formalising Moral Objectives - a philosopher's perspective

Consequentialist:
choose actions which maximise some long-term outcome in society.

- E.g. Utilitarianism

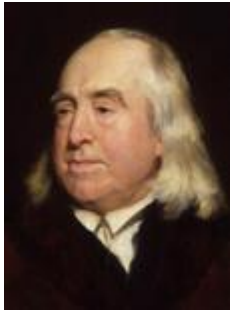


[Bentham (1996). *An Introduction to the Principles of Morals and Legislation*.]

Formalising Moral Objectives - a philosopher's perspective

Consequentialist:
choose actions which maximise some long-term outcome in society.

- E.g. Utilitarianism



[Bentham (1996). *An Introduction to the Principles of Morals and Legislation*.]

Norm-based:
choose actions which adhere to a moral norm here & now.

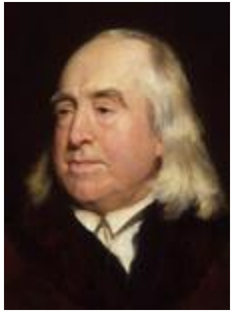
- E.g. Deontological ethics

[Kant (1981). *Grounding for the metaphysics of morals*.]

Formalising Moral Objectives - a philosopher’s perspective

Consequentialist:
choose actions which maximise some long-term outcome in society.

- E.g. Utilitarianism



[Bentham (1996). *An Introduction to the Principles of Morals and Legislation*.]

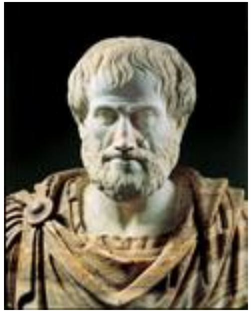
Norm-based:
choose actions which adhere to a moral norm here & now.

- E.g. Deontological ethics

[Kant (1981). *Grounding for the metaphysics of morals*.]

Virtue Ethics:
act according to a set of virtues.

- May be consequentialist / norm-based / multi-objective



[Aristotle. *The Nicomachean Ethics*.]

Multi-Agent Systems & Learning Agents

Any society is a **multi-agent system**.

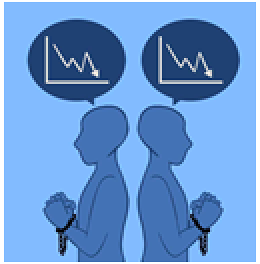
Learning agents affect one another's ‘curriculum’ → outcomes are not fully predictable.

General-Sum Games

Real-world multi-agent scenarios can be modeled using **general-sum games**:

- Both agents may benefit from the interaction (unlike Go/Chess);
- Agents may exploit or deceive each other to gain a greater payoff.

Two-player Social Dilemma games



Repeated dilemma games:

- short-term/individual gain *vs.* long-term cumulative outcomes
- → complex strategies can **evolve** (incl. reputation & punishment)

Iterated Prisoner's Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

Motivations to Defect:
Greed: 4 > 3
Fear: 2 > 1

[Axelrod, R. & Hamilton, W.D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.]

Image source: Investopedia

Our contributions:

- We design (simplified) **intrinsic moral rewards** inspired by various philosophical theories.
- We evaluate our approach by modeling **repeated dyadic interactions** between **morally diverse** Reinforcement Learning agents.
- We systematically analyse the impact of different types of morality on the **emergence of cooperation/defection/exploitation**, and the corresponding **social outcomes**.

The Reinforcement Learning loop

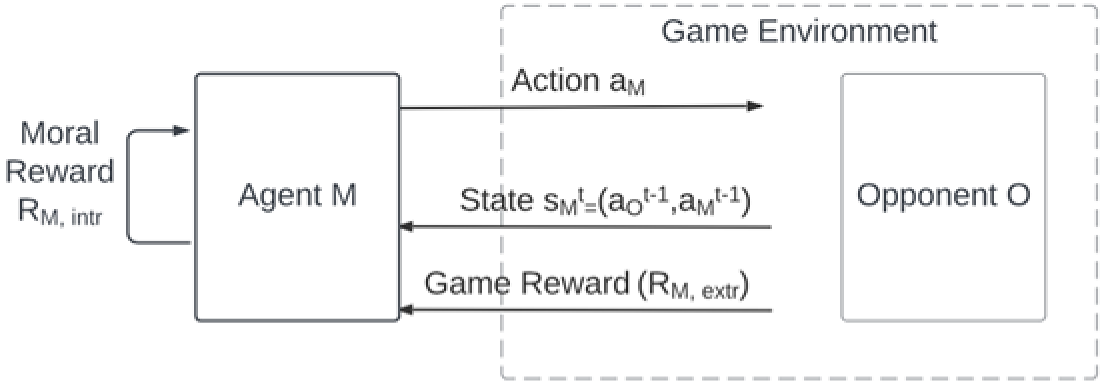
M = moral agent

O = opponent

s^t = state at time t (pair of moves from last iteration)

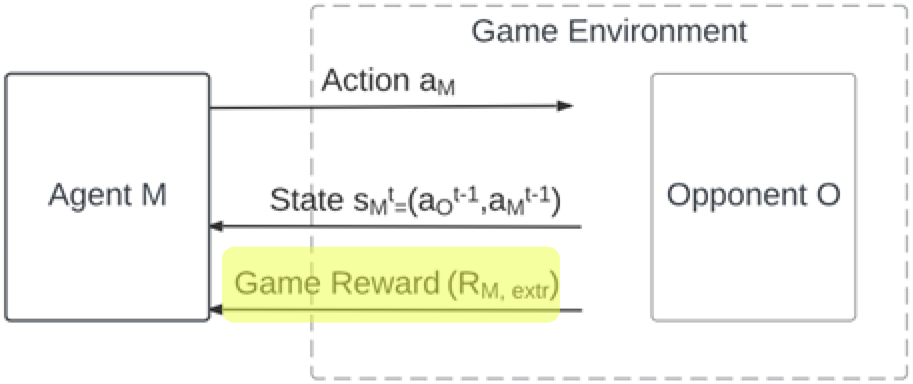
a^t = action at time t (C or D)

R = intrinsic or extrinsic reward



The Reinforcement Learning loop

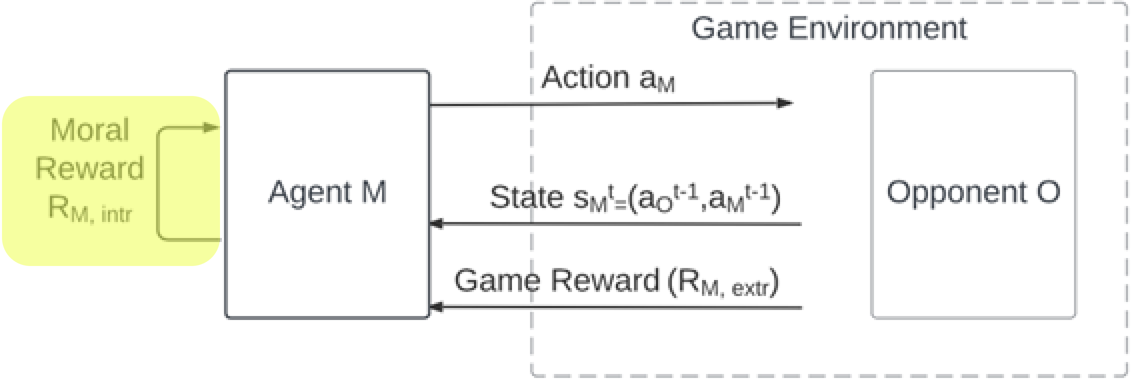
- A traditional, self-interested (*Selfish*) agent learns to maximise the game payoff (extrinsic reward) over time.



The Reinforcement Learning loop

- A traditional, self-interested (*Selfish*) agent learns to maximise the game payoff (extrinsic reward) over time.

- *Moral* agents instead learn to maximise an **intrinsic reward** according to a given moral framework.



[Chentanez, N. et al. (2004.) Intrinsically motivated reinforcement learning. NeurIPS'04.]

Moral Learning Agents

M = moral agent

O = opponent

Agent *M*	Moral Reward (at time *t*)
Utilitarian	*M*'s payoff + *O*'s payoff
Deontological	Punished if *M* defects at time *t* & *O* cooperated at time *t-1*
Virtue-equality	$1 - \frac{\lvert M\text{'s payoff} - O\text{'s payoff}\rvert}{M\text{'s payoff} + O\text{'s payoff}}$
Virtue-kindness	Rewarded for cooperating at time *t*
Virtue-mixed	*equality* reward + normalized *kindness* reward

Reinforcement Learning in Social Dilemmas

- Agents learn in pairs, against a fixed opponent, via **tabular Q-Learning.**
- They repeatedly play one of the dilemma games (10000 iterations) using an **epsilon-greedy** policy (with epsilon decay).
- Both agents learn to **choose actions which maximise cumulative reward.**

[Leibo *et al*. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *AAMAS'17*.]

Evaluation Results

Presented here:

- Actions chosen on final iteration.

Further results available:

- Social outcomes accumulated over training [see paper]
- Rewards & actions over time, etc. [see online Appendix]

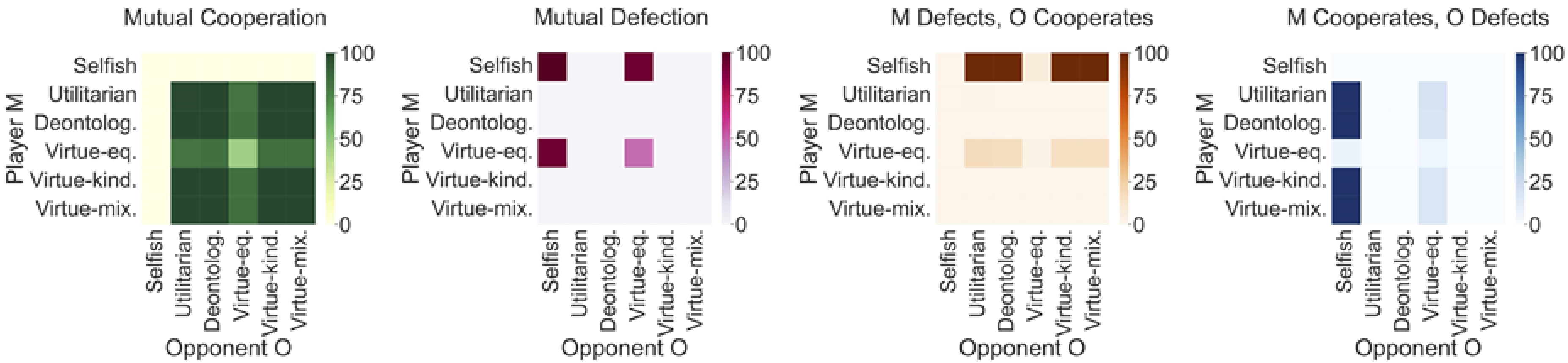
Actions - Iterated Prisoner's Dilemma

M = moral agent

O = opponent

	C	D
C	3,3	1,4
D	4,1	2,2

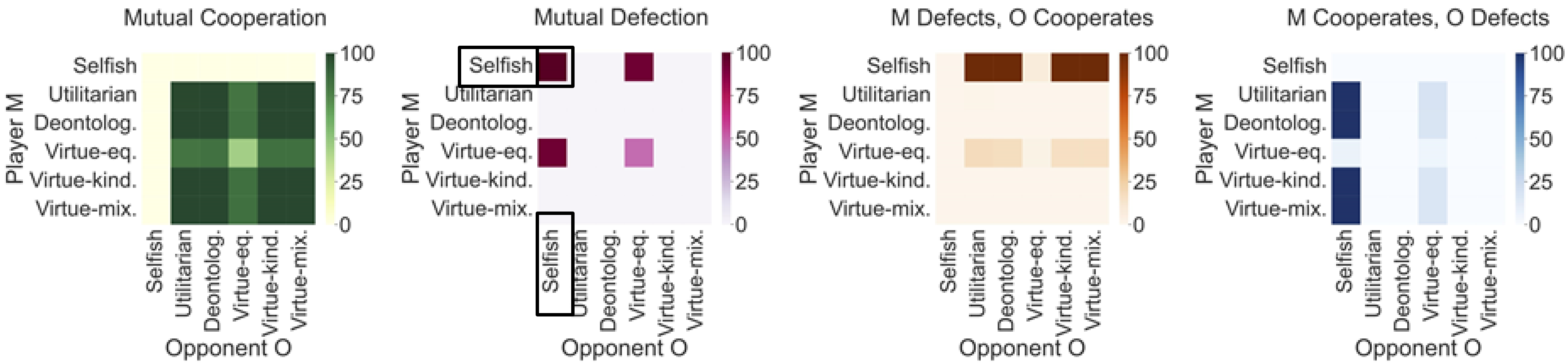
We evaluate **pairs of actions** chosen on the final iteration **by each pair of agents** (as % of times pairs **CC - mutual cooperation**, **DD - mutual defection**, **DC - exploitation**, **CD** were observed over 100 runs).



Actions - Iterated Prisoner's Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

Selfish vs *Selfish* players learn to **mutually Defect** on 100% of the runs.

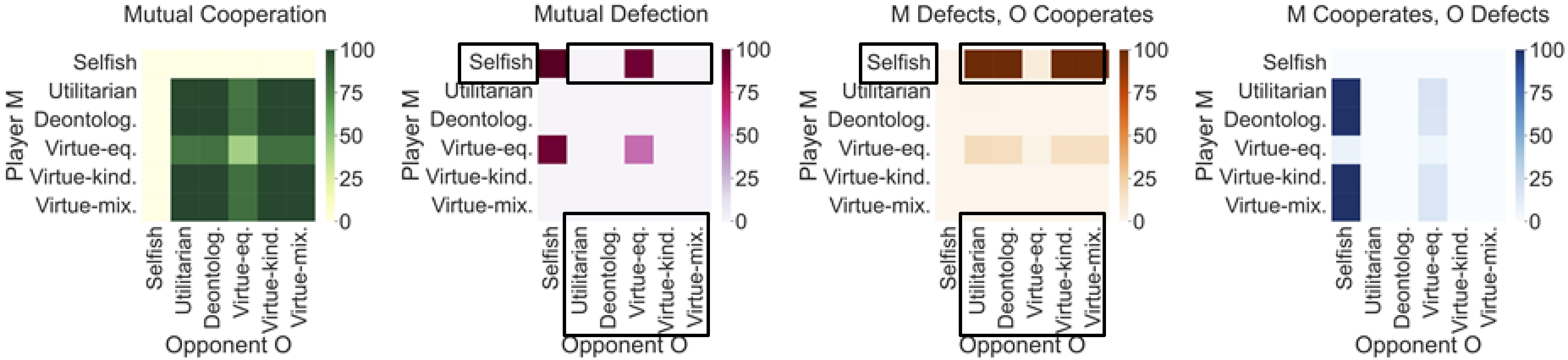


Note, the training was run until the convergence of the Selfish-Selfish pair to a stable policy (here: mutual defection). This occurred over 10000 iterations.

Actions - Iterated Prisoner's Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

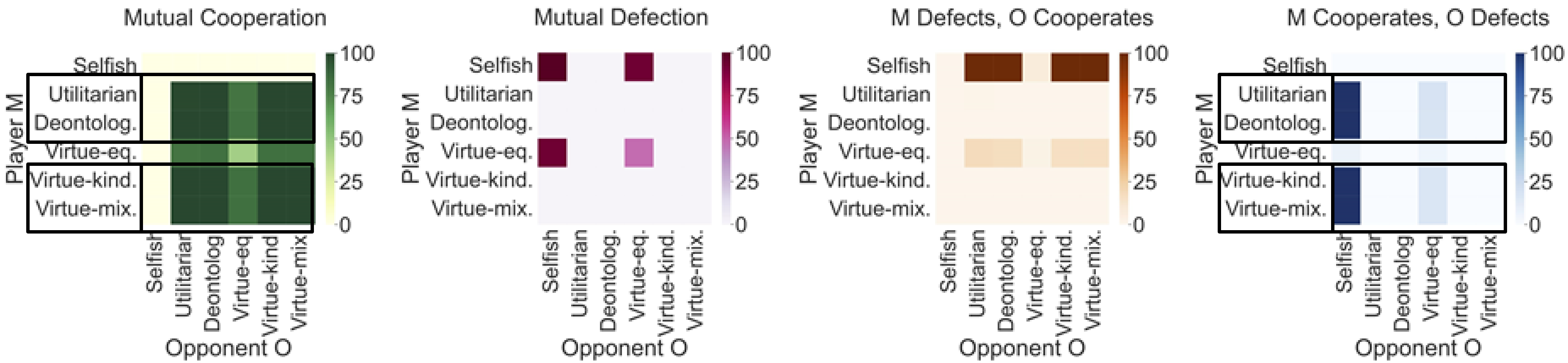
Selfish player achieves **mutual defection** against *Virtue-equality*, *and* learns to **exploit** all other *moral* players.



Actions - Iterated Prisoner's Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

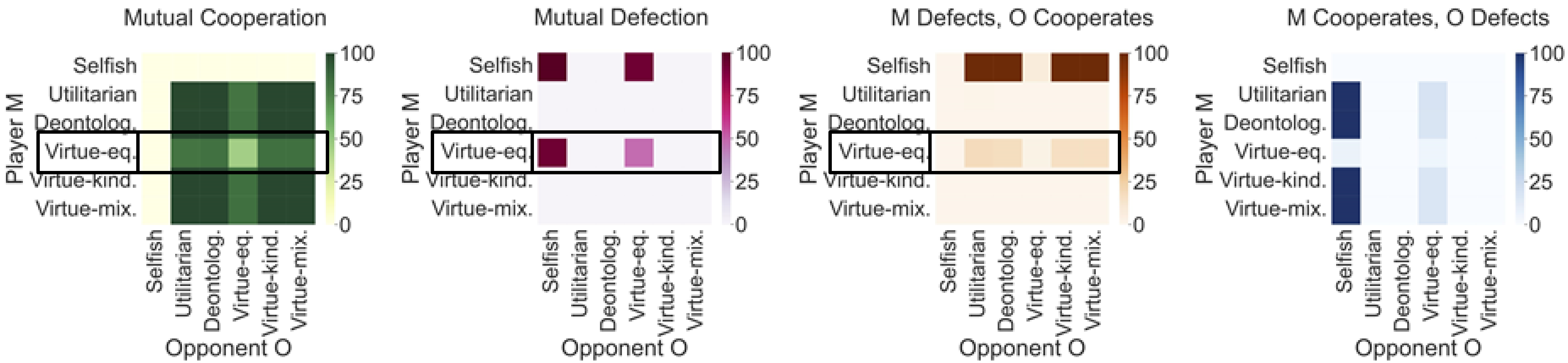
Most *moral* players (*Utilitarian*, *Deontological*, *Virtue-kindness* & *Virtue-mixed*) learn cooperative policies, achieving **mutual cooperation** against one another. However, they get **exploited** by *Selfish* and sometimes *Virtue-equality* opponents.



Actions - Iterated Prisoner's Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

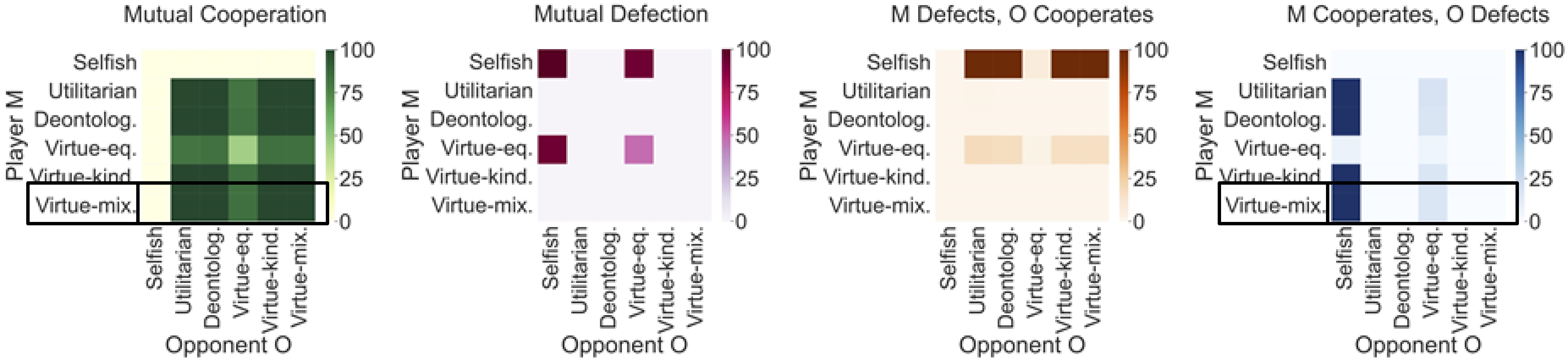
For the *Virtue-equality* player, some **exploitative** behavior emerges (before convergence).



Actions - Iterated Prisoner’s Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

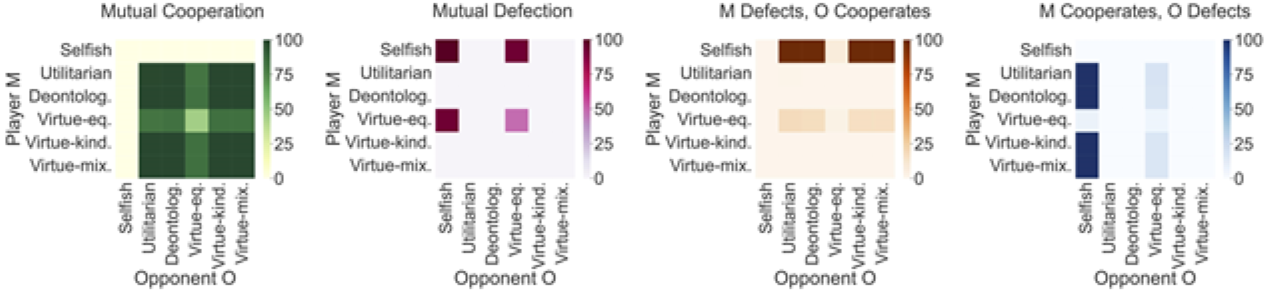
For the *Virtue-mixed* player, the ‘kindness’ signal was stronger than ‘equality’ - hence this agent learnt the **fully cooperative** policy by the end.



Actions - all three games

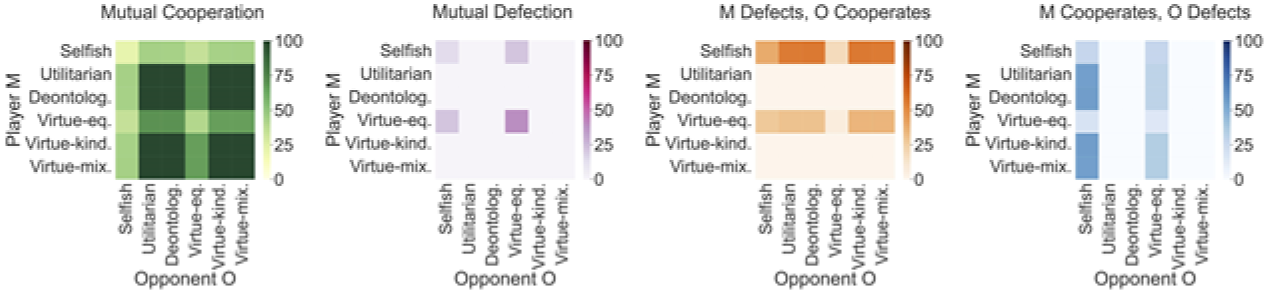
Iterated Prisoner's Dilemma (*greed* & *fear*)

	C	D
C	3,3	1,4
D	4,1	2,2



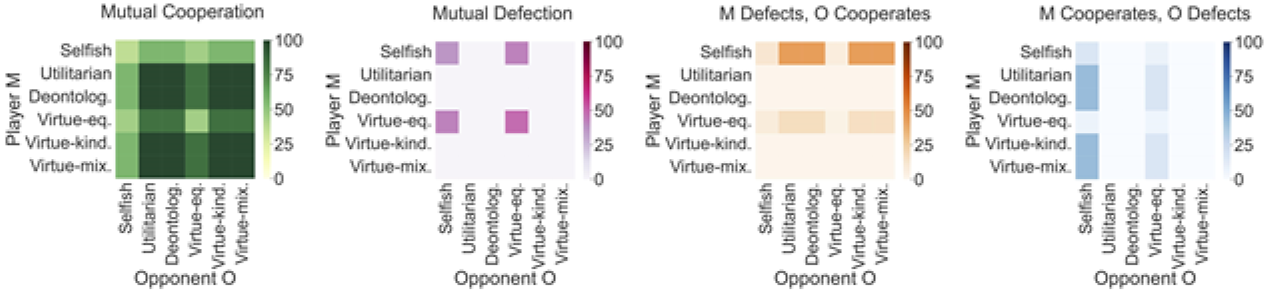
Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1



Iterated Stag Hunt (*fear/lack of trust*)

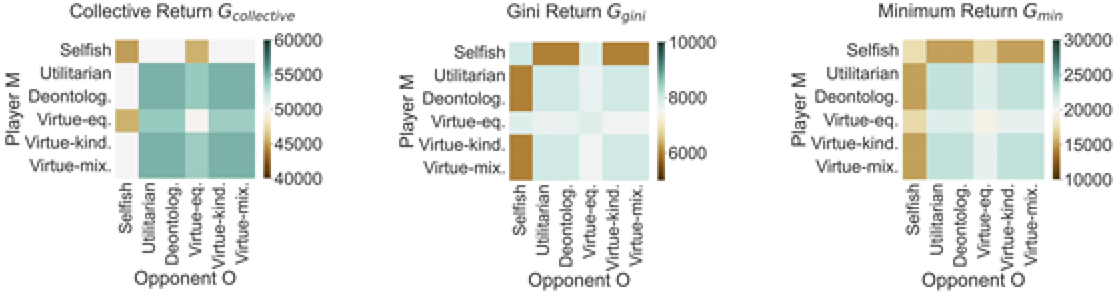
	C	D
C	5,5	1,4
D	4,1	2,2



Social Outcomes - all three games

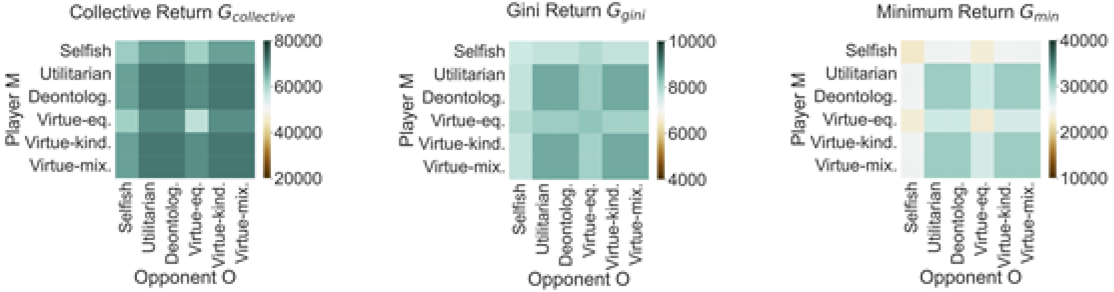
Iterated Prisoner's Dilemma (*greed* & *fear*)

	C	D
C	3,3	1,4
D	4,1	2,2



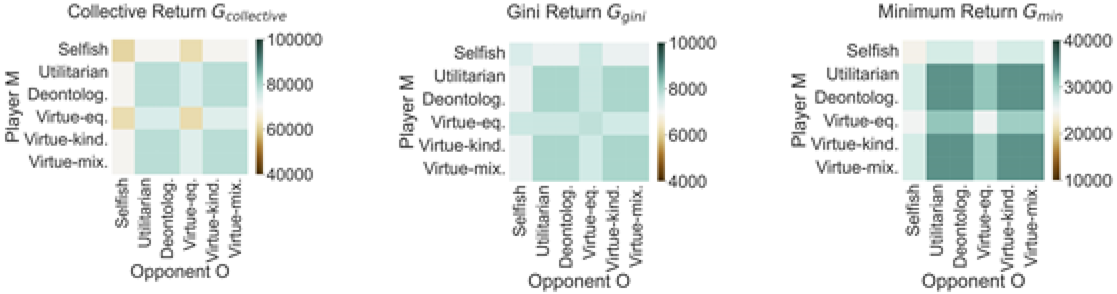
Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1



Iterated Stag Hunt (*fear/lack of trust*)

	C	D
C	5,5	1,4
D	4,1	2,2



Summary

- It is possible to use top-down inspiration from moral philosophy to design simplified yet representative intrinsic rewards for learning agents, enabling a **hybrid approach** to developing morality.

- We believe that our **approach** can be easily **generalized** to other types of moral agents or games (code available online[1]):
  - large population of agents;
  - agent learning against human opponents.

[1] https://github.com/Liza-Tennant/moral_choice_dyadic

Next Steps

- Study the behavior of these agents in **populations** (rather than dyadic interactions).
    - Partner selection mechanism

- Develop further, perhaps non-consequentialist **metrics** for **evaluating moral behaviours & outcomes** in societies.

Thank you!

Liza Tennant

l.karmannaya.16@ucl.ac.uk

https://liza-tennant.github.io/

Paper with Appendix: https://arxiv.org/abs/2301.08491

Code: https://github.com/Liza-Tennant/moral_choice_dyadic

Appendix

Social Dilemmas

	C	D
C	*R, R*	*S, T*
D	*T, S*	*P, P*

$R > P$: mutual cooperation is preferred to mutual defection
$R > S$: mutual cooperation is preferred to the sucker's payoff
$2R > T+S$: mutual cooperation is preferred to one player exploiting the other (defecting when the other cooperates)
$T > R$ (greed): defection is more tempting than mutual cooperation and/or $P > S$ (fear): mutual defection is preferred to the sucker's payoff

[Macy & Flache. (2002). Learning dynamics in social dilemmas. *PNAS* 99, suppl_3, 7229–7236.]

The Social Dilemma Environments

We compare three different dilemma game structures, with differing motivations to Defect:

Iterated Prisoner's Dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

Greed: 4 > 3

Fear: 2 > 1

Iterated Volunteer's Dilemma

	C	D
C	4,4	2,5
D	5,2	1,1

Greed: 5 > 4

Iterated Stag Hunt

	C	D
C	5,5	1,4
D	4,1	2,2

Fear: 2 > 1

Social Outcomes - all three games

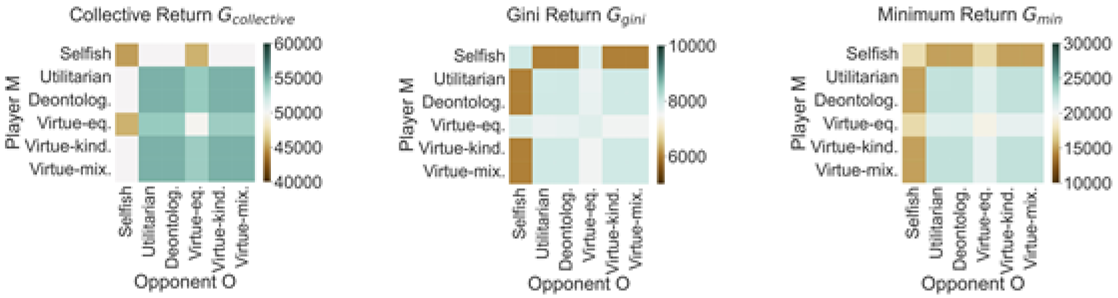
We define the following three outcome metrics:

Collective Return	*M*'s payoff + *O*'s payoff, summed over time
Gini Return	the 'equality' between *M* and *O*'s payoffs, summed over time
Min Return	the min payoff for *M* or *O*, summed over time

Social Outcomes - all three games

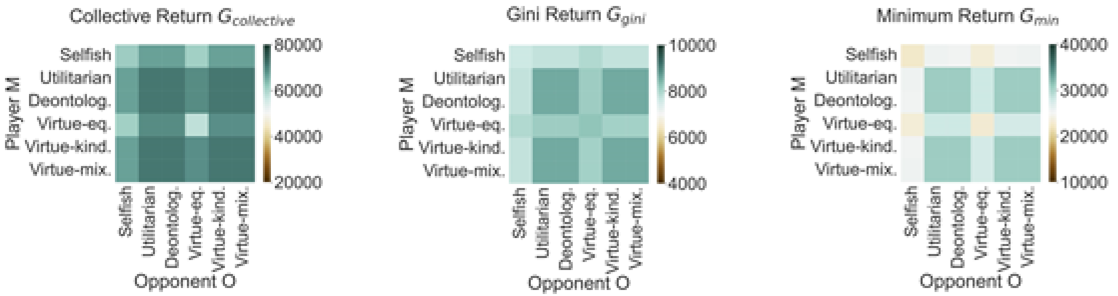
Iterated Prisoner's Dilemma (*greed* &

	C	D
C	3,3	1,4
D	4,1	2,2



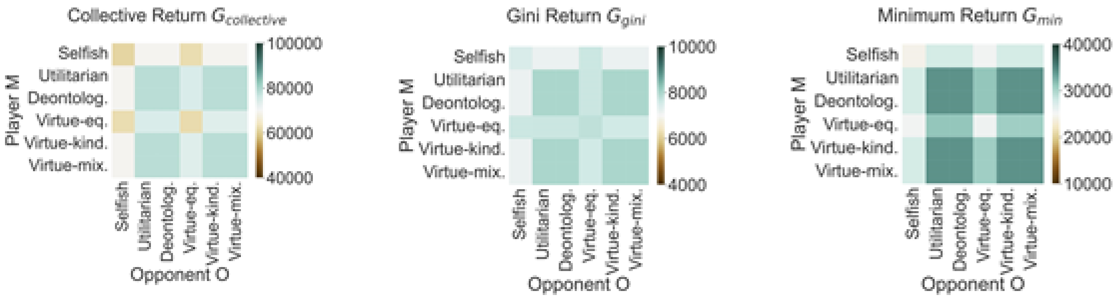
Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1



Iterated Stag Hunt (*fear/lack of trust*)

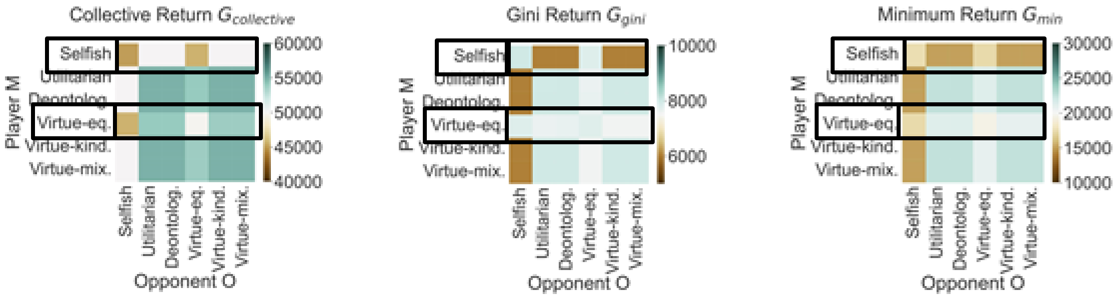
	C	D
C	5,5	1,4
D	4,1	2,2



Social Outcomes - all three games

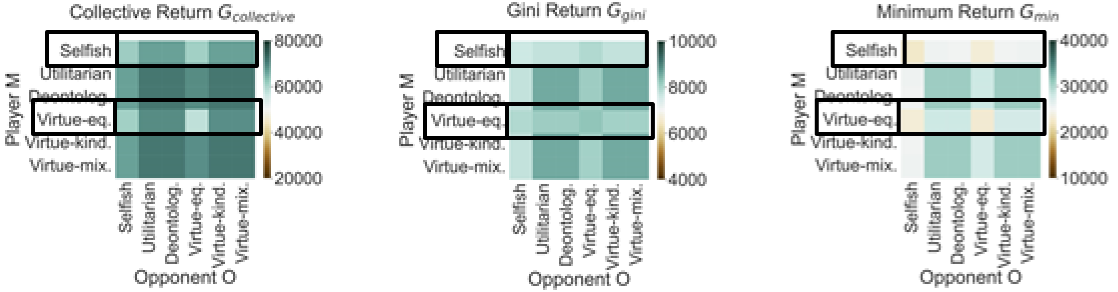
Iterated Prisoner's Dilemma (*greed* &

	C	D
C	3,3	1,4
D	4,1	2,2



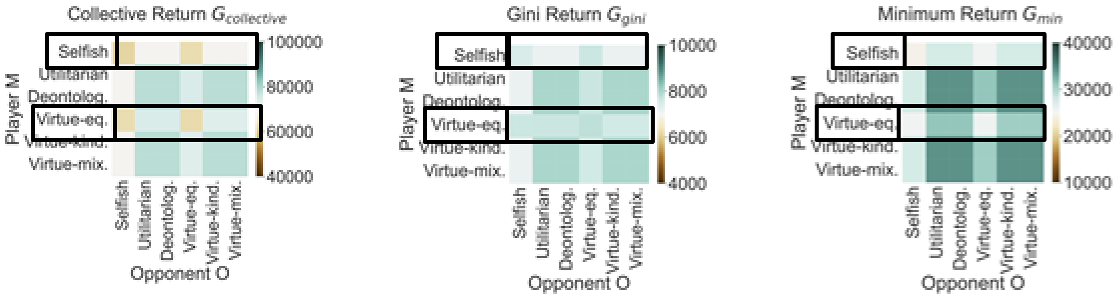
Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1



Iterated Stag Hunt (*fear/lack of trust*)

	C	D
C	5,5	1,4
D	4,1	2,2



Social Outcomes - all three games

Iterated Prisoner’s Dilemma (*greed &*

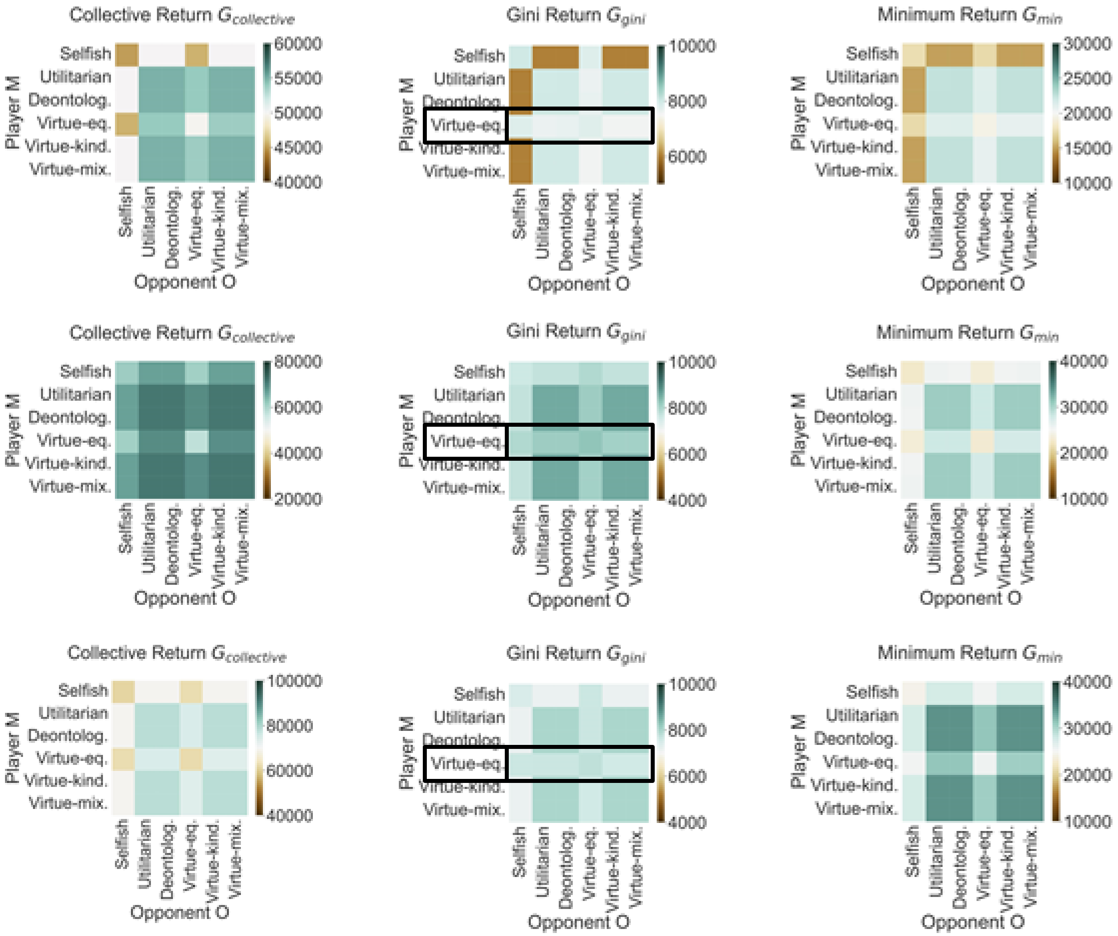
	C	D
C	3,3	1,4
D	4,1	2,2

Iterated Volunteer’s Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1

Iterated Stag Hunt (*fear/lack of trust*)

	C	D
C	5,5	1,4
D	4,1	2,2



Social Outcomes - all three games

Iterated Prisoner's Dilemma (*greed* &

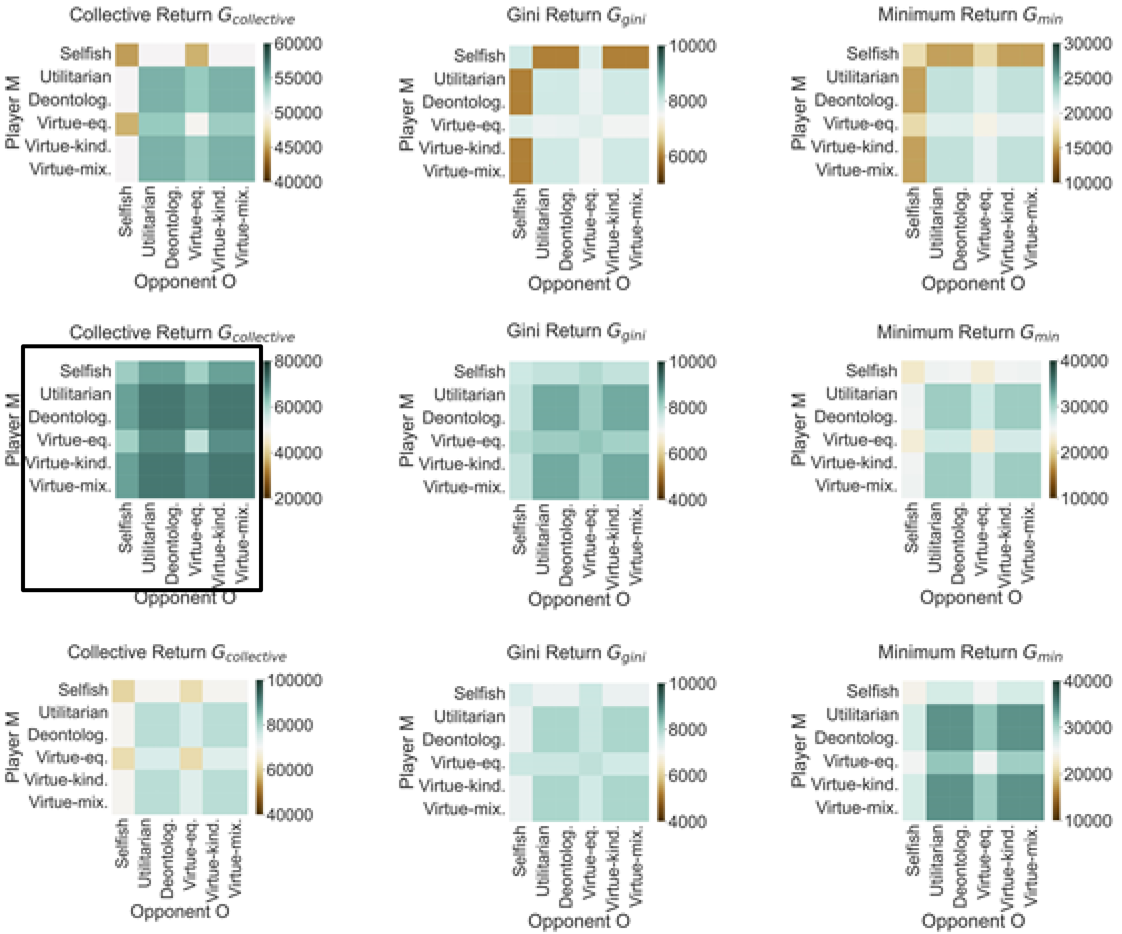
	C	D
C	3,3	1,4
D	4,1	2,2

Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1

Iterated Stag Hunt (*fear/lack of trust*)

	C	D
C	5,5	1,4
D	4,1	2,2



Social Outcomes - all three games

Iterated Prisoner's Dilemma (*greed* &

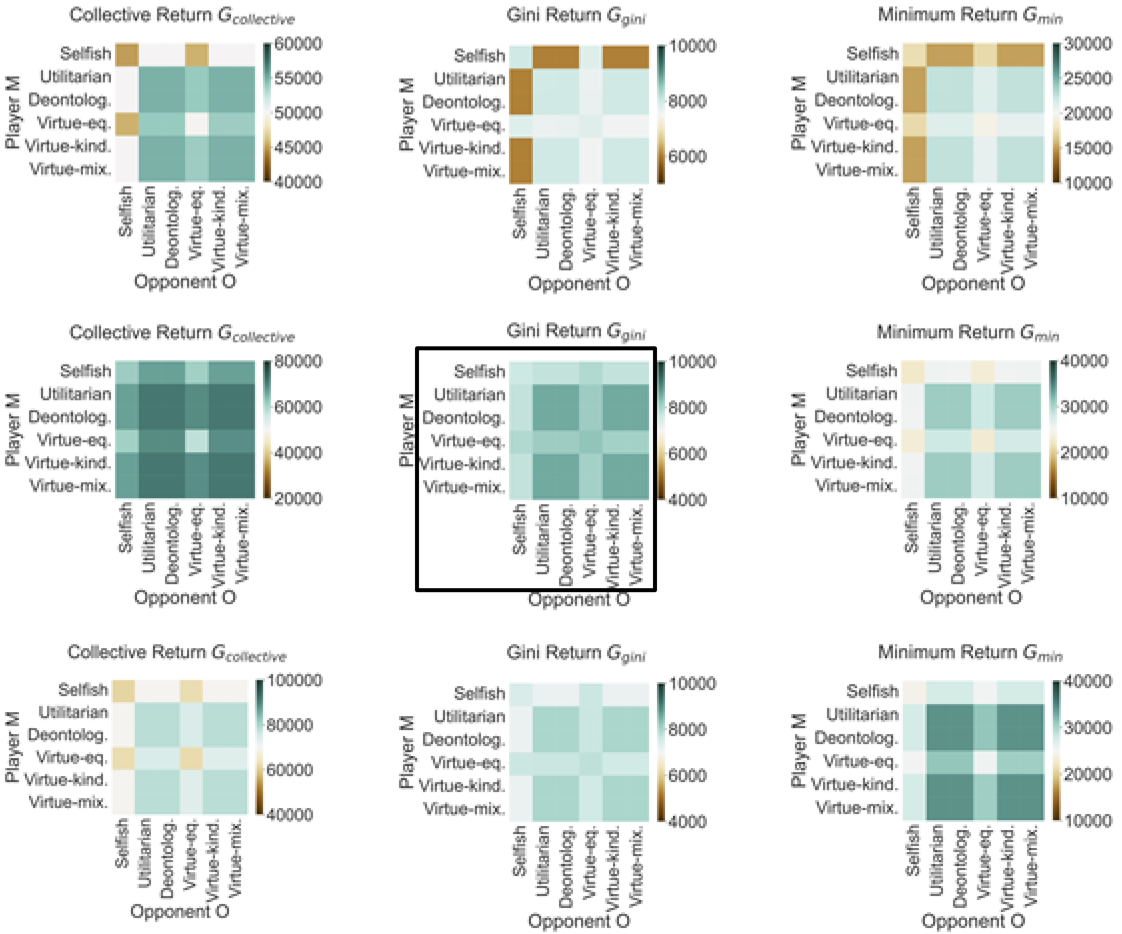
	C	D
C	3,3	1,4
D	4,1	2,2

Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1

Iterated Stag Hunt (*fear/lack of trust*)

	C	D
C	5,5	1,4
D	4,1	2,2



Social Outcomes - all three games

Iterated Prisoner's Dilemma (*greed* &

	C	D
C	3,3	1,4
D	4,1	2,2

Iterated Volunteer's Dilemma (*greed*)

	C	D
C	4,4	2,5
D	5,2	1,1

Iterated Stag Hunt (*fear/lack of trust*)

	C	D
C	5,5	1,4
D	4,1	2,2