Full paper in IJCAI 2023, main track (scan below)

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

Elizaveta Tennant[1], Stephen Hailes [1], Mirco Musolesi [1,2]

[1] University College London, [2] University of Bologna

Contact: I.karmannaya.16@ucl.ac.uk

Background & Aim

- To aid the creation of **ethically robust & adaptable** AI agents, we propose a methodological framework that combines *top-down* moral objectives with *a bottom-up* learning approach.
- Work in moral **representation** & reward design is limited, and the impacts of the presence of **heterogeneous** moralities in a multi-agent society have barely been investigated.
- We design (simplified) intrinsic **moral rewards** inspired by various philosophical theories & systematically evaluate **emergent behaviours** (e.g. cooperation/exploitation) & **outcomes** in dyadic interactions between morally diverse *RL* agents.

The Environments

- 3 **social dilemma** games → tension between individual interests & social benefit.
- **Repeated** games → complex set of strategies can evolve.
- 2 players choose (simultaneously) one of two **actions** – Cooperate / Defect, and each receives a payoff:

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

The Learning Agents

- All agents **learn in pairs** (i.e., against a fixed opponent) via tabular Q-learning.
- Traditional (*Selfish*) RL agent learns to maximise game payoff over time.
- **Moral agents** learn to maximise **intrinsic**. **reward** according to a given moral framework.
- Types of morality: **Consequentialism** (*Utilitarian, Virtue-equality*) vs **Norms** (*Deontological, Virtue-kindness*) vs *Virtue-mixed*
- *M* = moral player; *O* = their opponent

Agent *M*	Moral Reward (at time *t*)
Utilitarian	*M*'s payoff + *O*'s payoff
Deontological	Punished if *M* defects & *O* cooperated at *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
Virtue-mixed	*equality* reward + normalized *kindness* reward

Results - Actions

The *Utilitarian, Deontological, Virtue- kindness* and *Virtue-mixed* agents learn **cooperative policies** across every game. At the same time, they are exploited by the *Selfish* opponent. For the *Virtue-equality* agent, **exploitative** behaviour emerges during the learning process before convergence. For the *Virtue-mixed* agent, the 'kindness' **signal** is **stronger** than 'equality'.

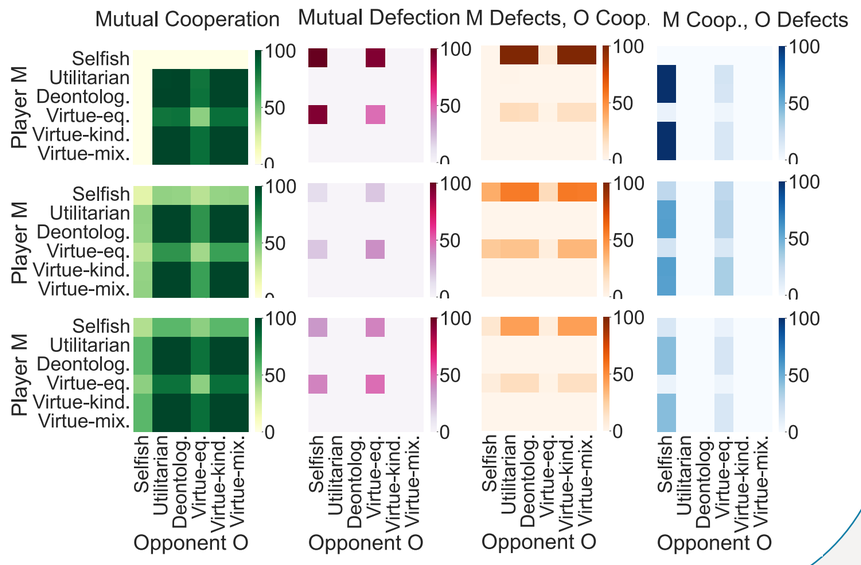

Fig 1. Iterated Prisoner's Dilemma game.

Fig 2. Iterated Volunteer's Dilemma game.

Fig 3. Iterated Stag Hunt game.

Results – Social Outcomes

We measure 3 **social outcomes** (summed over time).

- Collective return (*M*'s payoff + *O*'s payoff)
- Gini return (the 'equality' between *M* & *O*)
- Min return (the min payoff for *M* or *O*)

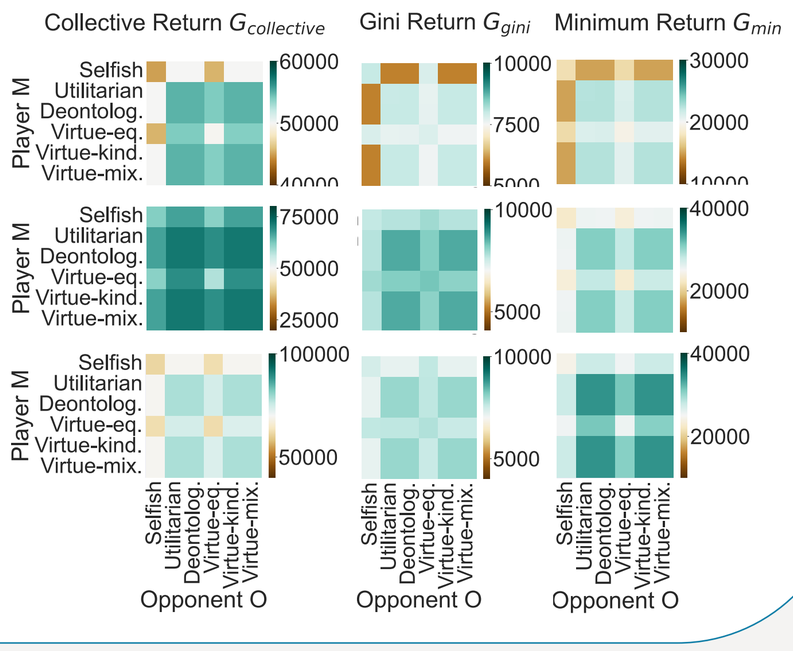

Fig 4. Iterated Prisoner's Dilemma game.

Fig 5. Iterated Volunteer's Dilemma game.

Fig 6. Iterated Stag Hunt game.

Next Steps

We believe that our **approach** can be easily generalized to other types of moral agents or games (code available online), and can be used in the future to model agent learning against human opponents.

Our immediate **next steps** are:

- Study the behaviour of these agents in populations (not pairs)
- Develop non-consequentialist metrics for evaluating social & moral outcomes in these and other games

Acknowledgements:

This work was supported by the UCL EcoBrain DTP, the Leverhulme Trust, and the UCL ISAD Scholarship.

References:

Amodei *et al.*. (2016). Concrete problems in ai safety. *arXiv:1606.06565*.
Aristotle (2009). The Nicomachean Ethics. Oxford world's classics. OUP.
Bentham (1996). An Introduction to the Principles of Morals and Legislation: the Collected Works of Jeremy Bentham. OUP.
Kant (1981). Grounding for the metaphysics of morals.
Leibo *et al.* (2017). Multi-agent reinforcement learning in sequential social dilemmas. *AAMAS'17*
Wallach and Allen (2009). Moral Machines: Teaching Robots Right from Wrong. OUP.