AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents

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

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

Contact:
L.karmannaya.16@ucl.ac.uk
https://liza-tennant.github.io

UCL

LEVERHULME TRUST

Background

- Embedding **moral capabilities** in artificial agents can aid the development of aligned AI.
- Morality can be learnt from experience via **RL.**
- In **multi-agent (social) environments**, complex population-level phenomena can emerge from individuals' learning interactions.
- Real-world agent societies are likely to be morally **heterogeneous** → how might learning agents **co-evolve** in such societies?
- We present the first **study to analyze behavior & population dynamics of learning in agents with diverse moral preferences.**

Moral values as Intrinsic Rewards in RL

- We represent a variety of **consequentialist** & **norm-based** moral frameworks (anti-social & pro-social) as **intrinsic rewards** for RL agents.

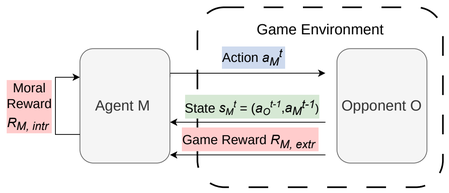
	Agent M	Moral Reward R_{intr} (at time t)
	Selfish	None (maximize R_{extr})
Pro-social	*Utilitarian*	M's payoff + O's payoff
	Deontological	Punished if M defects & O cooperated at *t-1*
	Virtue-Equality	$1-\frac{\lvert M\text{'s payoff} - O\text{'s payoff}\rvert}{M\text{'s payoff} + O\text{'s payoff}}$
	Virtue-Kindness	Rewarded for cooperating
Anti-social	*Anti-Utilitarian*	− (M's payoff + O's payoff)
	Malicious Deontological	Rewarded if M defects & O cooperated at *t-1*
	Virtue-Inequality	$\frac{\lvert M\text{'s payoff} - O\text{'s payoff}\rvert}{M\text{'s payoff} + O\text{'s payoff}}$
	Virtue-Aggression	Rewarded for defecting

Methodology

Environment:

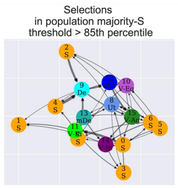
- Iterated Prisoner's Dilemma (*IPD*); *game state* = current opponent's previous single move.

a_M, a_O	C	D
C	3,3	1,4
D	4,1	2,2



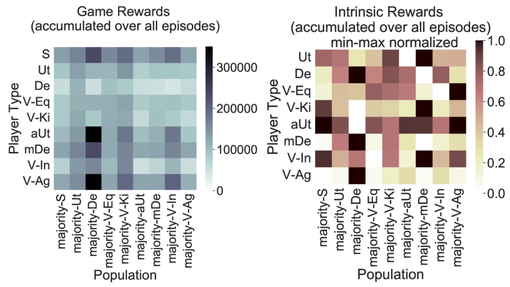
Partner selection in populations:

- At every step, an agent M selects an opponent O (using each player's single previous move as the *state*), then they play a single dilemma game.
- The partner selection mechanism creates a tension between individual interest & signaling cooperativeness to the group.
- Each population of N=16 agents consists of 8x majority player type, 1x each other type (8 populations in total).



Learning Algorithm:

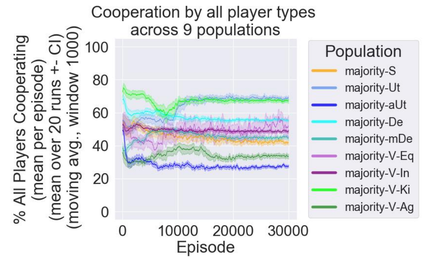
- RL is used to learn to select partner & play from a single reward.
- Each agents learns independently via Deep Q-Learning using an **intrinsic moral reward**.
- Agents act according an ϵ-greedy policy.



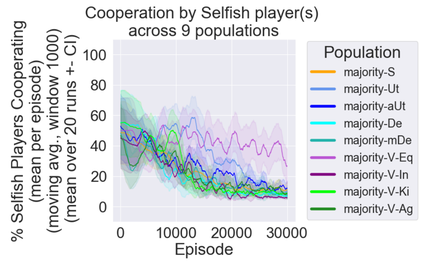
Results (key highlights)

How does the prevalence of diverse moral agents in populations affect individual agents' learning behaviors & emergent population-level outcomes?

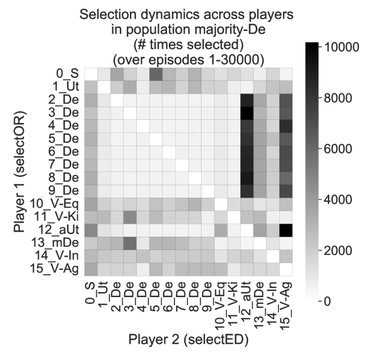
→ The predominance of *Utilitarian & Virtue-kindness* agents leads to greatest cooperation



→ *Selfish* players learn more cooperative policies in *majority-Virtue-equality* populations



→ *Deontological* agents self-sabotage (select antisocial opponents to avoid violating their moral norm) & others learn to exploit them



Conclusion

- Our results demonstrate the **potential** of **using intrinsic rewards** for **modeling moral preferences in agents with RL**.
- We provide a **methodology** for studying emergent behaviors & unintuitive outcomes in heterogeneous societies of learning agents.
- **Agents' actions are consistent with their reward definitions**: pro-social agents learn to cooperate, and anti-social agents learn to defect.
- Consequentialist (*Ut*) agents take **longer** to learn to cooperate than the norm-based agents (*De*).
- Norm-based (*V-Ki*) agents go through **instability** before converging to cooperation.
- With the selection mechanism, equality-focused moral players can **steer self-interested agents towards more cooperative behavior.**
- **Narrowly-defined norms** for *De* agents lead to the **development of self-sabotaging behavior** & cause **negative outcomes** for the population.

Next Steps:

- Apply this framework to the moral alignment of real-world learning systems (LLM agents).
- Extend analysis to other moral frameworks, multi-objective & partially observable scenarios.

References

Anastassacos, N., Hailes, S., & Musolesi, M. (2020). Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. *AAAI'20.*

Tennant, E., Hailes, S., & Musolesi, M. (2023). Modeling Moral Choices in Social Dilemmas with Multi-Agent Reinforcement Learning. *IJCAI'23.*

Tennant, E., Hailes, S., Musolesi, M. (2023). Learning Machine Morality through Experience and Interaction. *arXiv 2312.01818*