Navigating the Complexity of AI Cooperation: From Learning Dynamics to Language Models

Mariana Meireles[1] and Wolfram Barfuss[2]

Abstract. To ensure the alignment of advanced AI systems with human interests, AI needs social understanding and cooperative intelligence to integrate well into society. But what are the consequences if more intelligent actors (an advanced AI) are supposed to cooperate with less intelligent ones (an average human)? Here, we set out to study how differences in agents' intelligence impact collective cooperation by a combination of model simulations inspired by complex systems science and behavioral experiments conducted with large-language models.

1 Motivation

Artificial General Intelligence (AGI) refers to a type of artificial intelligence that can self-improve and thereby exceed human capabilities in solving individual tasks significantly. There's no guarantee that such advanced AI systems will have the same cooperative capabilities as humans (Dafoe et al., 2021). Humanity's collective intelligence is built upon the extraordinary ability of humans to cooperate (Bowles and Gintis, 2011). To ensure a desirable integration of advanced AI systems into human societies, it is, therefore, essential to improve our understanding of the cooperative capabilities of self-learning agents.

Here, we focus on how differences in agents' intelligence impact collective cooperation. In one plausible scenario, more intelligent agents have an advantage over less intelligent ones. They could exploit less intelligent agents to the disadvantage of those less intelligent agents. This scenario is highly undesirable and requires interventions in the further development of advanced AI systems. In another plausible scenario, when the goals of the intelligent agents depend on the less intelligent agents' actions, the more intelligent agent would facilitate cooperation for the benefit of both agents which is highly desirable.

There are many facets that can make one actor more intelligent than another. In this project, we focus on the internal representation of social information. We use a combination of model simulations inspired by complex systems science and behavioral experiments conducted with large-language models to investigate the conditions that make one or the other scenario more likely.

2 Methods

Learning dynamics. For the model simulations, we resort to the framework of multi-agent reinforcement learning (MARL). However, classical simulations of learning algorithms and advanced AI systems have significant disadvantages for improving the understanding of the learning agent's emergent collective behavior: they are noisy, sometimes hard to explain, sample-inefficient, and computationally intense. We, therefore, employ an approach to MARL inspired by complex system science and evolutionary game theory: collective reinforcement learning dynamics (CRLD) (Barfuss, 2022, 2023). CRLD is characterized by two types of idealization. First, CRLD uses dynamic learning equations as a lightweight model of the computationally intense reinforcement learning update. Second, CRLD aims to understand the principles behind emergent collective behavior in idealized, low-dimensional environments.

Environment. The problem of cooperation is most pronounced in social dilemmas. Here, individual incentives and collective welfare are not aligned. Individuals profit from exploiting others or fear being exploited by others, while at the same time, the collective welfare is maximized if all choose to cooperate (Dawes, 1980).

Here, we use the iterated (two-agent) Prisoner's dilemma as our test bed. Because of its simplicity in carving out the tensions between individual incentives and collective welfare, the effects discovered here are likely to exist also in higher-dimensional social dilemmas. Specifically, a cooperator is someone who pays a cost, c, for another individual to receive a benefit, b. A defector has no cost and does not deal out benefits (Nowak, 2006). Cost and benefit are measured in terms of the agents' utility (Schultz et al., 2017). Figure 1 shows the learning dynamics at the iterated Prisoner's dilemma when both agents learn a strategy which reacts to both agents' actions of the last round (i.e., CC,CD,DC,DD).

Extended Abstract.
[1]Impact Academy & BarfussLab, University of Bonn, GER; [2]University of Bonn, GER; | February 5, 2024

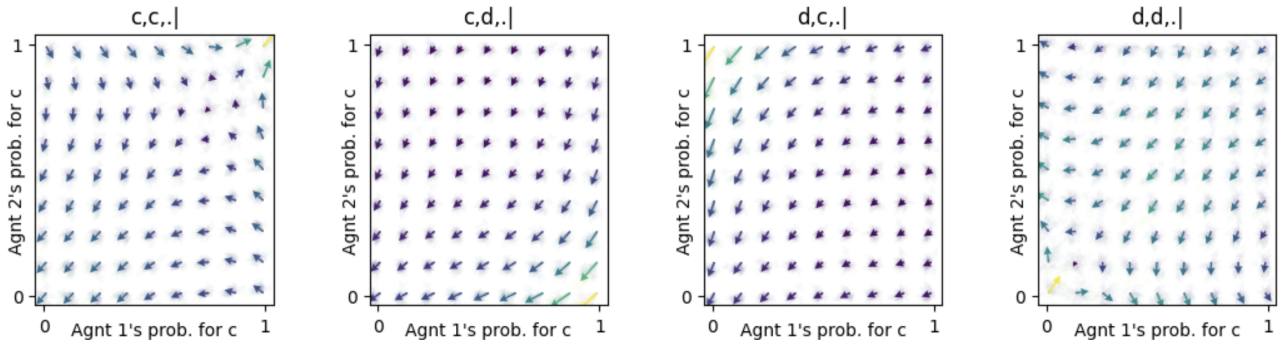

Figure 1: Learning dynamics in strategy phase space. There is a basin of attraction around mutual cooperation when both agents cooperated in the last round (CC), indicating the viability of overall cooperation even when benefits b exceed costs c (Barfuss and Meylahn, 2023). We will vary more and heterogeneous observation spaces.

LLM behavioral experiment. Besides the learning dynamic simulation, we investigate our scenarios also with behavioral experiments using large language models (LLMs). In order to avoid testing variance brought up by the intricacies of language, we use sentiment analysis tools in combination with tools from the field of survey methodology (Shaughnessy et al., 2000) to prompt these language models. By adopting this method, we were able to interact with the LLMs in a way that significantly reduced bias.

3 Conclusion

This marks the beginning of a three-month research project funded by the Impact Academy, whose results we are eager to discuss at the Machine+Behavior Conference 2024.

References

Wolfram Barfuss. Dynamical systems as a level of cognitive analysis of multi-agent learning. *Neural Computing and Applications*, 34(3):1653–1671, 2022. ISSN 1433-3058. doi: 10.1007/s00521-021-06117-0. URL https://doi.org/10.1007/s00521-021-06117-0.

Wolfram Barfuss. pyCRLD: Collective reinforcement learning dynamics in python, 2023. URL https://wbarfuss.github.io/pyCRLD/.

Wolfram Barfuss and Janusz M. Meylahn. Intrinsic fluctuations of reinforcement learning promote cooperation. *Scientific Reports*, 13(1):1309, 2023.

Samuel Bowles and Herbert Gintis. *A Cooperative Species*. Princeton University Press, Princeton, 2011. ISBN 9781400838837. doi: doi:10.1515/9781400838837. URL https://doi.org/10.1515/9781400838837.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

Robyn M Dawes. Social dilemmas. *Annual review of psychology*, 31(1):169–193, 1980.

Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

Wolfram Schultz, Wiliam R Stauffer, and Armin Lak. The phasic dopamine signal maturing: From reward via behavioural activation to formal economic utility. *Current Opinion in Neurobiology*, 43:139–148, 2017. ISSN 09594388. doi: 10.1016/j.conb.2017.03.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0959438817300892.

John J. Shaughnessy, Eugene B. Zechmeister, and Jeanne S. Zechmeister. *Research methods in psychology*. McGraw-Hill, 2000.