Investigating the Impact of Direct Punishment on the Emergence of Cooperation in Mulit-agent Reinforcement Learning Systems (Abstract Reprint)

Nayana Dasgupta[1], **Mirco Musolesi**[1,2]

[1]Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK
[2]Department of Computer Science and Engineering, University of Bologna, Via del Risorgimento 2, 40136, Bologna, Italy

Abstract Reprint. This is an abstract reprint of the journal article by Dasgupta and Musolesi (2025).

Abstract

Solving the problem of cooperation is fundamentally important for the creation and maintenance of functional societies. Problems of cooperation are omnipresent within human society, with examples ranging from navigating busy road junctions to negotiating treaties. As the use of AI becomes more pervasive throughout society, the need for socially intelligent agents capable of navigating these complex cooperative dilemmas is becoming increasingly evident. Direct punishment is a ubiquitous social mechanism that has been shown to foster the emergence of cooperation in both humans and non-humans. In the natural world, direct punishment is often strongly coupled with partner selection and reputation and used in conjunction with third-party punishment. The interactions between these mechanisms could potentially enhance the emergence of cooperation within populations. However, no previous work has evaluated the learning dynamics and outcomes emerging from multi-agent reinforcement learning populations that combine these mechanisms. This paper addresses this gap. It presents a comprehensive analysis and evaluation of the behaviors and learning dynamics associated with direct punishment, third-party punishment, partner selection, and reputation. Finally, we discuss the implications of using these mechanisms on the design of cooperative AI systems.

References

Dasgupta, N.; and Musolesi, M. 2025. Investigating the Impact of Direct Punishment on the Emergence of Cooperation in Mulit-agent Reinforcement Learning Systems. *Autonomous Agents and Multi-Agent Systems*, 39: 19.