Modelling Cooperation in Network Games with Spatio-Temporal Complexity

Extended Abstract

Michiel A. Bakker[1,2*], Richard Everett[1*], Laura Weidinger[1], Iason Gabriel[1], William S. Isaac[1], Joel Z. Leibo[1], & Edward Hughes[1]
1. DeepMind, 2. MIT
bakker@mit.edu,{reverett,lweidinger,iason,williamis,jzl,edwardhughes}@google.com

ABSTRACT

The real world is awash with multi-agent problems that require collective action by self-interested agents, from the routing of packets across a computer network [14] to the management of irrigation systems [10]. Such systems have local incentives for individuals, whose behavior has an impact on the global outcome for the group. Given appropriate mechanisms describing agent interaction, groups may achieve socially beneficial outcomes, even in the face of short-term selfish incentives. In many cases, collective action problems possess an underlying graph structure, whose topology crucially determines the relationship between local decisions and emergent global effects. Such scenarios have received great attention through the lens of network games. However, this abstraction typically collapses important dimensions, such as geometry and time, relevant to the design of mechanisms promoting cooperation. In parallel work, multi-agent deep reinforcement learning has shown great promise in modelling the emergence of self-organized cooperation in complex gridworld domains [5, 7, 11]. Here we apply this paradigm in graph-structured collective action problems.

KEYWORDS

Multi-Agent Reinforcement Learning, Cooperation, Network Games

ACM Reference Format:
Michiel A. Bakker[1,2*], Richard Everett[1*], Laura Weidinger[1], Iason Gabriel[1], William S. Isaac[1],, Joel Z. Leibo[1], & Edward Hughes[1]. 2021. Modelling Cooperation in Network Games with Spatio-Temporal Complexity: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 OUR CONTRIBUTIONS

The aims of this paper are (1) to introduce a new method for modelling the behavior of self-interested agents in collective action problems with topological, geometrical and temporal structure, and (2) to use this method to draw conclusions relevant for mechanism design that promotes cooperative behavior. Our method comprises a protocol for intervening on a spatio-temporally complex environment, and modelling the effects on social outcomes for rational agents via multi-agent deep reinforcement learning. To illustrate this general method, we introduce a new gridworld domain called *Supply Chain*, in which agents are rewarded for processing goods according to a given network structure. For environmental interventions, we take the perspective of a system designer and ask: what mechanisms might we introduce to the world, and how do these affect the cooperation of agents? We not only vary the topology of the world, as in traditional network games, but also the geometry, maintenance cost, and agent specialization. In all cases we find an intricate interplay between incentive structure, multi-agent interaction and learnability affecting the nature of emergent cooperation. More precisely, we introduce a metric of care in order to understand these dynamics [13]. We find that reciprocal care is diminished when the maintenance burden is lower, and that reciprocity is promoted by training generalist agents that can operate any station in the supply chain, rather than specialists. We do not expect the conclusions we draw to have general applicability; rather we argue that this case study demonstrates the power and insight provided by our new method.

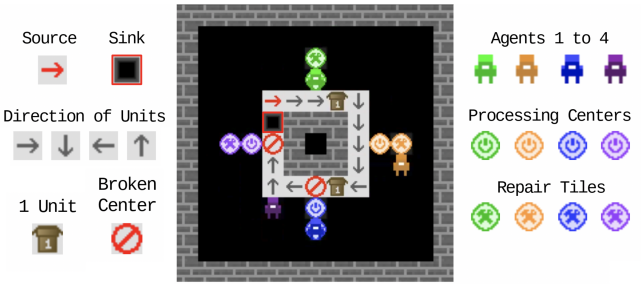


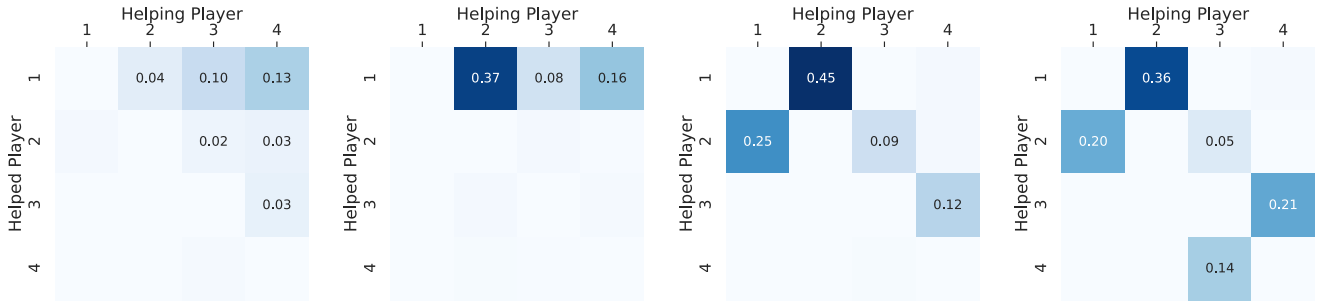
Figure 1: The Supply Chain environment, visualized with a circular layout mid-episode.

2 METHODS

The Supply Chain environment (Figure 1) is a 2-dimensional gridworld in which agents must maintain their own individual *processing centers* with the help of other agents, in order to process *units* passing through the supply chain. Units enter the environment via the source tile and are removed at the the sink tile. Importantly, units stop next to each processing center and do not continue along the supply chain until they have been processed by that processing center's owner. This is achieved by the agent standing on their processing tile, thereby processing the unit allowing it to continue and giving +1 reward to that agent. Upon processing a unit, there is a 25% chance that the agent's processing center breaks down which stops units from passing through the supply chain. The center can

*these authors contributed equally.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.



(a) After $\approx 2 \cdot 10^7$ training steps. **(b) After $\approx 4 \cdot 10^7$ training steps.** **(c) After $\approx 1.2 \cdot 10^8$ training steps.** **(d) After $\approx 1.45 \cdot 10^8$ training steps.**

Figure 2: Care matrices at the end of the four distinct learning phases. For improved readability, values below 0.01 are omitted.

be repaired by two agents standing on both the processing tile and associated repair tile. In the full paper we also present experiments that explore delayed automatic repair [2]. An episode lasts 1000 steps. On each step, there is a 10% chance of a unit entering the environment, and all other units in the supply chain move once if the next space in the supply chain is unoccupied. If the next space is occupied by another unit, the moving unit is discarded from the supply chain and cannot be processed by any agent, leading to a lost opportunity to obtain reward. Collective action is required to maintain the processing centers. Each agent would prefer that others took on the responsibility for fixing broken processing centers, since this comes at the opportunity cost of processing units themselves. However, if all agents refuse to cooperate, they receive low group reward. Depending on the environment properties we see different patterns of "care" between agents, understood in terms of help provided to others when repairing broken processing centers.

Consider an instance of the *Supply Chain* game with episode length T and N agents, uniquely assigned to N processing centers such that agent i always processes units at center i. Let R^i, B^i, and C^{ij}, be integer variables that specify, aggregated over one episode, how many times agent i respectively processes a unit, breaks its processing center, or repairs a processing center of agent j. The *care matrix* (C) with elements C^{ij} tracks the care (repairs) each agent has received from each other agent, relative to the total number of breakages $\sum_i B^i$. We introduce a more elaborate set of social outcome metrics in the full version of the manuscript [2].

We train agents using advantage actor-critic (A3C) [8] with 400 parallel environments to generate experience for each learner. At each step, agents receive 13x13 RGB pixels as state information. Episodes contain 4 agents which are sampled without replacement from a population of 8 and assigned to random processing centers in the environment [6]. Every agent uses their own neural network and is trained for 10^9 steps by receiving importance-weighted policy updates [4]. For additional details, please refer to the full paper [2].

3 THE EMERGENCE OF CARE

In this section, we analyze the learning dynamics of a single run in the environment with a circular layout (Figure 1). In the full version of the paper [2], we study, at convergence, the effect of changing if and how fast processing centers can repair themselves autonomously, we increase the inter-center distance to study the influence of geometry, we compare specialist with generalist agents, and we discuss how subtle changes in the environment's underlying graph structure can drastically change social outcomes.

Learning to care for others only benefits agents indirectly which makes it a more complex behavior to learn than processing units. Learning happens in distinct phases, each characterized by different behaviors and social outcomes. Each phase features a rather abrupt change in the individual reward received: once an agent has found an improved strategy via exploration, it is quickly able to exploit this, shifting the equilibrium dramatically. Accordingly, each phase can last shorter or longer depending on when the agents "discover" the new behavior. We therefore analyze the outcomes after each phase, similarly to [1, 11], and analyze one archetypical run. Detailed social outcome metrics, averages over multiple runs, and a second typical run can be found in the full version [2].

In Figure 2 we find the care matrix after each phase. Phase 1 begins at the start and is characterized by agents learning how to navigate the environment, process units and explore repairing. Agents have yet to learn when repairing is beneficial, and we thus find the highest average reward for the agent that is closest to the source. Phase 2 is characterized by a drastic increase in care. The second agent (and to a lesser extent the third and fourth) learn that they can earn more reward by repairing the first processing center. In phase 3, agent 2 learns that repairing center 1 only results in more reward when it can process units at its own center. To keep this incentive for agent 2, agent 1 thus learns to reciprocate the care received by agent 2 resulting in a drastic increase in reward for agent 2. Finally, during phase 4, reciprocity emerges between agents 3 and 4 as agent 3 learns that agent 4 only repairs when its own processing center is fixed. In a sense, it is surprising that reciprocity emerges in this environment. Naïvely, one might expect that the collective action problem is dominated by selfish incentives, and an explicit model-based intervention may be required to solve the social dilemma, as in [3, 12]. However, in our case, the underlying graph structure organizes agent interactions in such a way as to promote the emergence of reciprocity. This is exactly in line with previous work in abstract network games [9], extending it to a setting where we can examine mechanism design in detail.

Acknowledgements We would like to thank Theophane Weber, Kevin McKee and many other colleagues at DeepMind for useful discussions and feedback on this work.

REFERENCES

[1] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. arXiv:1909.07528 [cs.LG]

[2] Michiel A Bakker, Richard Everett, Laura Weidinger, Gabriel Iason, William S. Isaac, Joel Z. Leibo, and Edward Hughes. 2021. Modelling Cooperation in Network Games with Spatio-Temporal Complexity. arXiv:2102.06911 [cs.MA]

[3] Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z. Leibo. 2019. Learning Reciprocity in Complex Sequential Social Dilemmas. arXiv:1903.08082 [cs.MA]

[4] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. 2018. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *International Conference on Machine Learning*. 1407–1416.

[5] Ernst Fehr and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114, 3 (1999), 817–868. http://www.jstor.org/stable/2586885

[6] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 364, 6443 (2019), 859–865.

[7] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 464–473.

[8] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[9] Hisashi Ohtsuki and Martin A Nowak. 2007. Direct reciprocity on graphs. *Journal of theoretical biology* 247, 3 (2007), 462–470.

[10] Elinor Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807763

[11] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*. 3643–3652.

[12] Alexander Peysakhovich and Adam Lerer. 2017. Consequentialist conditional cooperation in social dilemmas with imperfect information. (10 2017).

[13] Joan C. Tronto. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. Routledge.

[14] David H Wolpert and Kagan Tumer. 2002. Collective intelligence, data routing and braess' paradox. *Journal of Artificial Intelligence Research* 16 (2002), 359–387.