The Imitation Game: Learned Reciprocity in Markov games

Extended Abstract

Tom Eccles
DeepMind
London, UK
eccles@google.com

Edward Hughes
DeepMind
London, UK
edwardhughes@google.com

János Kramár
DeepMind
London, UK
janosk@google.com

Steven Wheelwright
DeepMind
London, UK
sjwheel@google.com

Joel Z. Leibo
DeepMind
London, UK
jzl@google.com

ABSTRACT

Reciprocity is an important feature of human social interaction and underpins our cooperative nature. What is more, simple forms of reciprocity have proved remarkably resilient in matrix game social dilemmas. Most famously, the tit-for-tat strategy performs very well in tournaments of Prisoner's Dilemma. Unfortunately this strategy is not readily applicable to the real world, in which options to cooperate or defect are temporally and spatially extended. Here, we present a general online reinforcement learning algorithm that displays reciprocal behavior towards its co-players. We show that it can induce pro-social outcomes for the wider group when learning alongside selfish agents, both in a 2-player Markov game, and in 5-player intertemporal social dilemmas. We analyse the resulting policies to show that the reciprocating agents are strongly influenced by their co-players' behavior.

ACM Reference Format:
Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z. Leibo. 2019. The Imitation Game: Learned Reciprocity in Markov games. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Sustained cooperation among multiple individuals is a hallmark of human social behavior, and may even underpin the evolution of our intelligence [7, 17]. Often, individuals must sacrifice some personal benefit for the long-term good of the group, for example to manage a common fishery or provide clean air. Logically, it seems that such problems are insoluble without the imposition of some extrinsic incentive structure [12]. Nevertheless, small-scale societies show a remarkable aptitude for self-organization to resolve public goods and common pool resource dilemmas [13]. Reciprocity provides a key mechanism for the emergence of collective action, since it rewards for pro-social behavior and punishes for anti-social acts. Indeed, it is a common norm shared by diverse societies [1, 2, 14, 19]. Moreover, laboratory studies find experimental evidence for conditional cooperation in public goods games; see for example [5].

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

By far the most well-known model of reciprocity is Rapoport's Tit-for-Tat [16], for playing the repeated Prisoner's Dilemma game against an unknown opponent. The algorithm cooperates on its first move, and thereafter mimics the previous move of its partner, by definition displaying perfect reciprocity. Although Tit-for-Tat and its variants have proved resilient to modifications in the matrix game setup [3, 6, 11], it is clearly not applicable to realistic situations. In general, cooperating and defecting require an agent to carry out complex sequences of actions across time and space, and the payoffs defining the social dilemma may be delayed. In this setting, agents must learn both the high-level strategy of reciprocity and the low level policies required for implementing (gradations of) cooperative behavior.

Previous approaches [8, 10, 15] propose reinforcement learning models for 2-agent problems, based on a planning approach. Our approach differs in that it is model-free, making it practically applicable to more complex environments, and is able to reciprocate to a range of behaviors, rather than switching between pre-determined policies. It also does not rely on observing the rewards of other players.

2 MODEL

We propose an online-learning model of reciprocity which can be applied to complex social dilemmas. Our setup comprises two types of reinforcement learning agents, *innovators* and *imitators*. An innovator optimizes for a purely selfish reward. An imitator has two components: (1) a mechanism for measuring the level of sociality of different behaviors and (2) an intrinsic motivation [4] for matching the sociality of others.

We investigate two mechanisms for assessing sociality. The first is based on hand-crafted features of the environment. The other uses a learned "niceness network", which estimates the effect of one agent's actions on another agent's future returns, hence providing a measure of social impact [9]. More precisely, this network is trained to estimate the expected return for the imitator given the innovator's state and action, and a baseline given only the innovator's state. The difference between these is an estimate of how much the innovator's action advantaged the imitator. This is what we use to model the niceness of a single action. The quantity which the imitator is rewarded for imitating is a time-weighted sum of the niceness of the actions in the agent's trajectory.

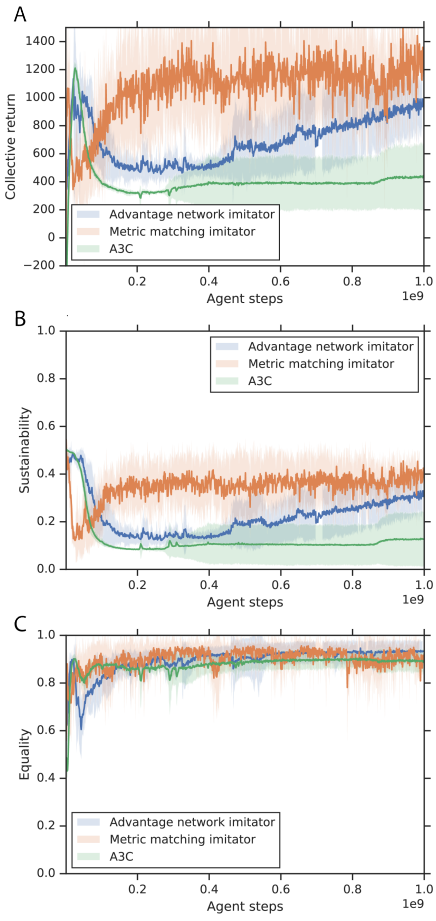

Figure 1: The effect of reciprocity on social outcomes in Harvest. (A) Collective return is higher when metric-matching or niceness-network imitators co-learn with a selfish innovator. (B) In the imitation conditions, the group learns a more sustainable strategy. (C) Equality remains high throughout training, suggesting that the imitators are successfully matching the cooperativeness of innovators.

3 EXPERIMENTS

Our main experiments have one innovator agent, learning alongside one or more imitator agents. The key hypothesis is that the innovator will learn to behave pro-socially. This is because the imitators are reciprocating towards them on a short timescale; this means that pro-social behaviour by the innovator leads to pro-social behaviour by all agents, and so to good outcomes for all agents.

We run this experiment in three environments. The first is Coins, a 2-player environment introduced in [10]. This environment has simple mechanics, and a strong social dilemma between the two players, similar to the Prisoner's Dilemma. This allows us to study our algorithms in a setup close to the Prisoner's Dilemma, and make comparisons to previous work. The other two environments are Harvest and Cleanup. These are more complex environments, with delayed results of actions, partial observability of a somewhat complex gridworld, and more than two players. These environments are designed to test the main hypothesis of this paper, that our algorithms are able to learn to reciprocate in complex environments where reciprocity is temporally extended and hard to define. We choose these two environments because they represent different classes of social dilemma; Cleanup is a public goods game, while Harvest is a common pool resource dilemma.

4 RESULTS

In all environments, we find that the outcomes of the groups including imitators are better for all agents than the outcomes for selfish agents in the same environment, both for agents matching hard-coded and learned metrics. This supports the hypothesis that our imitators are able to learn to reciprocate, and so induce pro-social behaviour in their selfish co-players. We also see other evidence for this – measures of prosociality and equality of returns are high and well-matched between the imitators and innovators. In Figure 1, we show these results for the Harvest environment.

To analyse the behaviour of the system, we measure the influence the trained agents have on each others' policies using techniques from [18]. This shows that the influence of the innovators on the imitators is much higher than for other pairs of agents. We also perform an ablation study, which shows that the imitation of niceness is the crucial component in our imitator model.

In the Coins game, we find that our models are not able to elicit the perfect cooperation seen from planning models in the same environments [10]. We believe this is because the reciprocity from the model-free algorithm is not as clear; this leaves open an important question of how to learn models of reciprocation which are both clear and scalable to complex environments.

5 CONCLUSION

Our reciprocating agents demonstrate an ability to elicit cooperation in otherwise selfish individuals, both in 2-player and 5-player social dilemmas. This reciprocation improves social outcomes for the whole group, with all agents contributing to the social good. Our algorithm scales well to complex environments, as it does not rely on planning.

REFERENCES

[1] L.C. Becker. 1990. *Reciprocity.* University of Chicago Press. https://books.google.co.uk/books?id=dWgI4lI7h-cC
[2] P.M. Blau. 1964. *Exchange and Power in Social Life.* J. Wiley. https://books.google.co.uk/books?id=qhOMLscX-ZYC
[3] Robert Boyd. 1989. Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology* 136, 1 (1989), 47 – 56. https://doi.org/10.1016/S0022-5193(89)80188-2
[4] Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. 2005. Intrinsically Motivated Reinforcement Learning. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou (Eds.). MIT Press, 1281–1288. http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf
[5] Rachel Croson, Enrique Fatas, and Tibor Neugebauer. 2005. Reciprocity, matching and conditional cooperation in two public goods games. *Economics Letters* 87, 1 (2005), 95 – 101. https://doi.org/10.1016/j.econlet.2004.10.007
[6] Peter Duersch, Joerg Oechssler, and Burkhard C. Schipper. 2013. When is tit-for-tat unbeatable? *CoRR* abs/1301.5683 (2013). arXiv:1301.5683 http://arxiv.org/abs/1301.5683
[7] R. I. M. Dunbar. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* 16, 4 (1993), 681âĂŞ694. https://doi.org/10.1017/S0140525X00032325
[8] Max Kleiman-Weiner, Mark K Ho, Joe L Austerweil, Littman Michael L, and Joshua B. Tenenbaum. 2016. Coordinate to cooperate or compete: abstract goals

and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

[9] Bibb Latané. 1981. The psychology of social impact. *American Psychologist* 36(4) (1981). http://dx.doi.org/10.1037/0003-066X.36.4.343

[10] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *CoRR* abs/1707.01068 (2017). arXiv:1707.01068 http://arxiv.org/abs/1707.01068

[11] M. A. Nowak. 2006. *Evolutionary Dynamics*. Harvard University Press. https://books.google.co.uk/books?id=YXrIRDuAbE0C

[12] Mancur Olson. 1965. *The Logic of Collective Action*. Harvard University Press. https://books.google.co.uk/books?id=jzTeOLtf7_wC

[13] E. Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press. https://books.google.co.uk/books?id=4xg6oUobMz4C

[14] Elinor Ostrom. 1998. A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association, 1997. *American Political Science Review* 92, 1 (1998), 1âĂŞ22. https://doi.org/10.2307/2585925

[15] Alexander Peysakhovich and Adam Lerer. 2017. Consequentialist conditional cooperation in social dilemmas with imperfect information. *CoRR* abs/1710.06975 (2017). arXiv:1710.06975 http://arxiv.org/abs/1710.06975

[16] A. Rapoport, A.M. Chammah, and C.J. Orwant. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press. https://books.google.co.uk/books?id=yPtNnKjXaj4C

[17] Simon M. Reader and Kevin N. Laland. 2002. Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences* 99, 7 (2002), 4436–4441. https://doi.org/10.1073/pnas.062041299 arXiv:http://www.pnas.org/content/99/7/4436.full.pdf

[18] Andrea Tacchetti, H Francis Song, Pedro AM Mediano, Vinicius Zambaldi, Neil C Rabinowitz, Thore Graepel, Matthew Botvinick, and Peter W Battaglia. 2018. Relational Forward Models for Multi-Agent Learning. *arXiv preprint arXiv:1809.11044* (2018).

[19] J.W.A. THIBAUT and H.H. Kelley. 1966. *The Social Psychology of Groups*. Wiley. https://books.google.co.uk/books?id=KDH5Hc9F2AkC