game can end stochastically at any time period and also ends when any player drops the ball.

Each of these games has, at a high level, the basic Stag Hunt property that there exists a strategy which guarantees a safe payoff and a risky strategy which only works if one's partner commits to it. However, unlike in the matrix Stag Hunt, these strategies are no longer single labeled actions, but rather complex policies which map the state of the world to an action to be taken. See the full paper for a more in depth description of the games as well as parameters that we vary in our experiments.

We compare the performance (here in terms of payoff to our agent) from situations where both agents are selfish, both agents are prosocial, and where only our agent is prosocial. In all conditions, both agents start with randomly initialized policies and learn via deep RL by playing with each other (see full paper for deep RL training details). Figure 1 shows a sample of our main results: the intuition from the matrix game replicates in these more complex environments. Giving just a single agent social preferences can help lead both agents to coordinate on payoff-dominant strategies in these more complex Stag Hunt-like games.

Other aspects of the game play important roles in setting the potential benefits and costs of choosing a prosocial strategy and we discuss these at length in the full paper (available on arxiv). We also discuss extending our main results to the case of Stag Hunt games played on simple networks. In addition, we discuss the relationship between prosociality and other types of learning modifications that have been proposed in the literature. These include optimism in the form of lenient learning [10, 12] or Frequency Maximum Q-learning [9] and potential-based reward shaping [1, 4, 5].

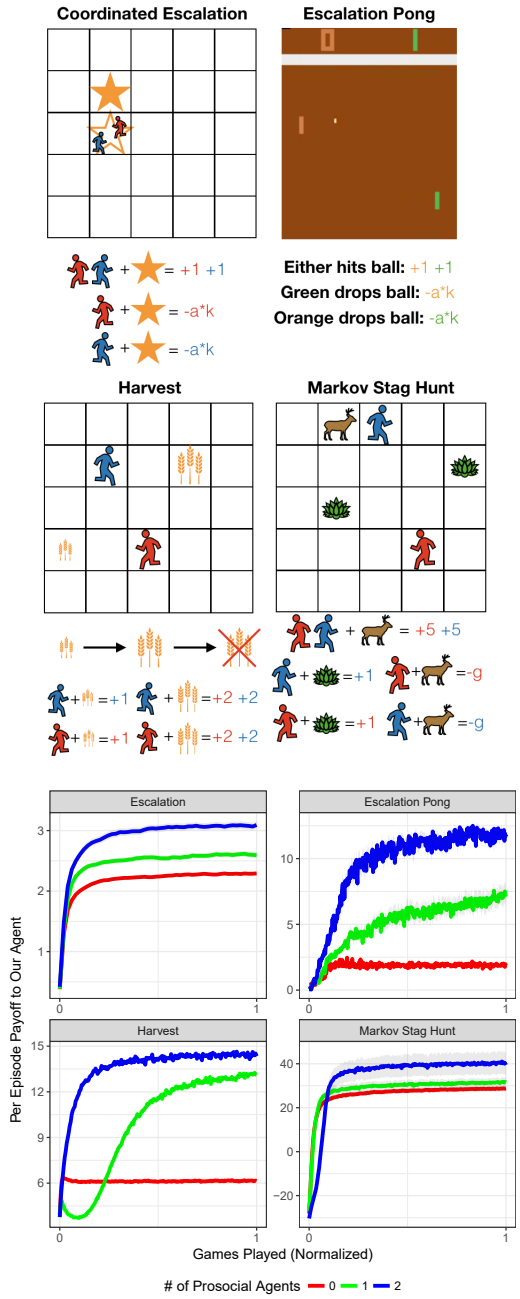

Figure 1: The intuitions from the 2×2 Stag Hunt generalize to more complex Markov games. Lines reflect average payoffs over replicates smoothed over 1000 episode blocks. Error bars reflect standard errors estimated using independent replicates.

REFERENCES

[1] Monica Babes, Enrique Munoz De Cote, and Michael L Littman. 2008. Social reward shaping in the prisoner's dilemma. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 1389–1392.

[2] Colin Camerer. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

[3] Hans Carlsson and Eric Van Damme. 1993. Global games and equilibrium selection. *Econometrica: Journal of the Econometric Society* (1993), 989–1018.

[4] Sam Devlin and Daniel Kudenko. 2011. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 225–232.

[5] Sam Devlin, Daniel Kudenko, and Marek Grześ. 2011. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems* 14, 02 (2011), 251–278.

[6] Drew Fudenberg and David K Levine. 1998. *The theory of learning in games*. Vol. 2. MIT press.

[7] John C Harsanyi, Reinhard Selten, et al. 1988. A general theory of equilibrium selection in games. *MIT Press Books* 1 (1988).

[8] Michihiro Kandori, George J Mailath, and Rafael Rob. 1993. Learning, mutation, and long run equilibria in games. *Econometrica: Journal of the Econometric Society* (1993), 29–56.

[9] Spiros Kapetanakis and Daniel Kudenko. 2002. Reinforcement learning of coordination in cooperative multi-agent systems. *AAAI/IAAI* 2002 (2002), 326–331.

[10] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. 2012. Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *The Knowledge Engineering Review* 27, 1 (2012), 1–31.

[11] Martin A Nowak. 2006. *Evolutionary dynamics*. Harvard University Press.

[12] Liviu Panait, Karl Tuyls, and Sean Luke. 2008. Theoretical advantages of lenient learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research* 9, Mar (2008), 423–457.

[13] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one* 12, 4 (2017), e0172395.

[14] John B Van Huyck, Raymond C Battalio, and Richard O Beil. 1990. Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review* 80, 1 (1990), 234–248.

[15] Wako Yoshida, Ray J Dolan, and Karl J Friston. 2008. Game theory of mind. *PLoS computational biology* 4, 12 (2008), e1000254.