Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones

Extended Abstract

Alexander Peysakhovich*
Facebook AI Research
alexpeys@fb.com

Adam Lerer*
Facebook AI Research
alerer@fb.com

KEYWORDS

coordination; game theory; reinforcement learning

ACM Reference Format:
Alexander Peysakhovich* and Adam Lerer. 2018. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 2 pages.

1 EXTENDED ABSTRACT

Real world interactions are full of coordination problems [2, 3, 8, 14, 15] and thus constructing agents that can solve them is an important problem for artificial intelligence research. One of the simplest, most heavily studied coordination problems is the matrix-form, two-player Stag Hunt. In the Stag Hunt, each player makes a choice between a risky action (hunt the stag) and a safe action (forage for mushrooms). Foraging for mushrooms always yields a safe payoff while hunting yields a high payoff if the other player also hunts but a very low payoff if one shows up to hunt alone. This game has two important Nash equilibria: either both players show up to hunt (this is called the payoff dominant equilibrium) or both players stay home and forage (this is called the risk-dominant equilibrium [7]).

In the Stag Hunt, when the payoff to hunting alone is sufficiently low, dyads of learners as well as evolving populations converge to the risk-dominant (safe) equilibrium [6, 8, 10, 11]. The intuition here is that even a slight amount of doubt about whether one's partner will show up causes an agent to choose the safe action. This in turn causes partners to be less likely to hunt in the future and the system trends to the inefficient equilibrium.

We are interested in the problem of agent design: our task is to construct an agent that will go into an initially poorly understood environment and make decisions. Our agent must learn from its experiences to update its policy and maximize some scalar reward. However, there will also be other agents which we do not control. These agents will also learn from their experiences. We ask: if the environment has Stag Hunt-like properties, can we make changes to our agent's learning to improve its outcomes? We focus on reinforcement learning (RL), however, many of our results should generalize to other learning algorithms.

In this paper, we show that adding prosociality - making our agent get reward from others receiving reward - is a simple strategy for improving coordination. In the full version of this paper we show show analytically that in matrix Stag Hunt we can improve the probability of dyads of agents converging to the good equilibrium even if we can only make a single agent prosocial. We experimentally generalize these insights in a domain where analytical solutions are difficult: Markov games with Stag Hunt-like structure and learning via function approximation (deep reinforcement learning).

We consider 3 different grid-world games: Markov Stag Hunt, Harvest and Escalation (Figure 1). In each of these games two agents move on a 5×5 grid and can move in any of the 4 cardinal directions. In the Markov Stag Hunt the grid is populated with a Stag and 2 plants. Moving over a plant gives either agent 1 point and causes the plant to disappear and re-appear in another part of the board. Moving over the Stag causes an agent to lose g points but if both agents move over the stag simultaneously they each gain 5 points and the stag disappears and re-appears randomly elsewhere on the board. In each time period the stag moves towards the closest agent to it, although the stag can never catch an agent who continues to move away from it.

In the Harvest game at each time step a plant can appear randomly somewhere on the board (up to 4 plants can be on the board at a time). Each plant is born small, becomes mature and then dies. Players can move over plants to pick them up. Players receive 1 point if they up a young plant, however waiting until each plant becomes mature and picking it up yields 2 points to *both* players.

In the Coordinated Escalation game a special marker appears on one of the squares. If the agents step on the square together, they both receive one point, at which point an adjacent square lights up. If the agents step together onto the next square, they receive 1 point. If at any time an agent breaks the streak (eg. by stepping off the path), the other agent receives a penalty of some multiplier times the current length of the streak, and the game ends. The current streak length T is observed (encoded in the state). This game has many equilibria, where both agents play to keep streaks of size T but no more with risk escalating at each time step (as the reward from further escalation is always 1 but the cost from one's partner failing to continue the pattern increases linearly).

We also consider a version of Escalation where agents must learn from raw pixels. We use methods employed in [13] to adapt Atari Pong to construct Escalation Pong. In Escalation Pong each agent controls a player, each time the ball is hit both agents receives a reward of 1, however if an agent drops the ball (allows it to pass) then *the other agent* receives a reward of $-k$ where k is proportional to the number of times the ball has been hit back and forth. The

*Both authors contributed equally to this paper. Author order is random. The full version of this paper is available online: https://arxiv.org/abs/1709.02865

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

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].

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.

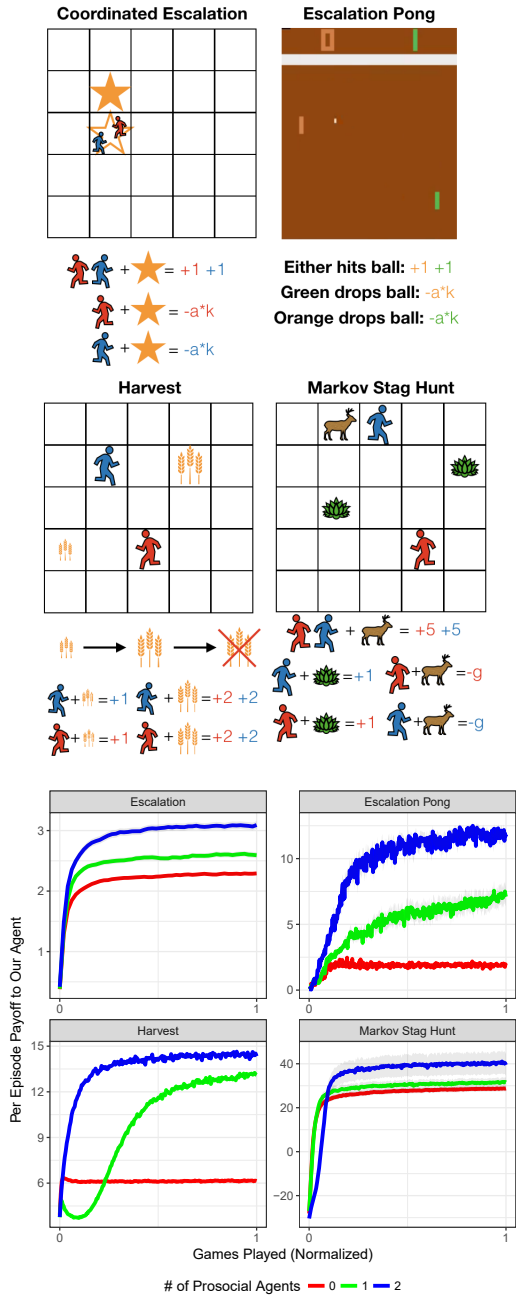

Figure 1: The intuitions from the 2×2 Stag Hunt generalize to more complex Markov games. Lines reflect average payoffs over replicates smoothed over 1000 episode blocks. Error bars reflect standard errors estimated using independent replicates.

[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.