Learning Robust Helpful Behaviors in Two-Player Cooperative Atari Environments

Extended Abstract

Paul Tylkin
Harvard University
ptylkin@g.harvard.edu

Goran Radanovic
Max Planck Institute for Software Systems
gradanovic@mpi-sws.org

David C. Parkes
Harvard University
parkes@eecs.harvard.edu

ABSTRACT

We study the problem of learning helpful behavior, specifically, learning to cooperate with differently-skilled and diverse partners in the context of two-player, cooperative Atari games. We show robust performance of these so-called *Helper-AIs* when paired with different kinds of partners (both human and artificial agents), including partners that they have not previously encountered during training. In particular, while pairing an expert AI with a non-expert AI leads to performance that is worse than when pairing the non-expert AI with a copy of itself, these Helper-AIs provide a substantial boost in joint performance.

KEYWORDS

Reinforcement Learning; Multi-Agent Systems; Human-AI Collaboration

ACM Reference Format:
Paul Tylkin, Goran Radanovic, and David C. Parkes. 2021. Learning Robust Helpful Behaviors in Two-Player Cooperative Atari Environments: 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 INTRODUCTION

We advance the study of cooperative behavior through applied research in the domain of two-player Atari games. There is already a rich tradition of using Atari to drive advances in AI [10, 11]. Atari games are designed to be fun for people to play, challenging enough to test AI methods, and have been well-studied in single-player settings (e.g., [17]). While Atari games can support multi-player modes (e.g., [16]), to the best of our knowledge, there have only been a few works that use Atari as a test-bed for studying cooperative behavior [8, 15]. These papers mainly focus on emergent behavior and social dilemmas in AI-AI interaction in Pong.

In our research, we are studying AI-AI and AI-human cooperation in the context of two-player Atari games that have richer game dynamics than Pong. In this extended abstract, we present results for two-player Space Invaders, modified here to make it a cooperative environment. In particular, we configure the game so that players maximize the joint score and to remove the bonus for loss of life of the other player.

Our main interest is to understand whether reinforcement learning can be used to achieve *helpful behavior*— where one agent is trained to follow a policy that will help a second, *partner agent*. We want to understand whether this can be done in a robust way: *can the same helpful AI cooperate effectively with a diverse set of partners, both artificial and human?*

For our human-subject experiments, we recruit participants on Amazon Mechanical Turk and make use of a new web-based framework, introduced in this paper. All experiments are conducted subject to oversight by Harvard's IRB. The *Javatari Learning Environment* (JLE) framework allows human subjects to interact with AIs through in-browser Atari emulators. The JLE framework makes use of a modified version of *Javatari* [13] to support in-browser Atari play by humans, enabling easy crowdsourcing of game trajectories.

We use the *ACKTR* [17] algorithm for reinforcement learning (as provided as a part of *OpenAI Baselines* [5]), together with *OpenAI Gym* [2] and the *Arcade Learning Environment (ALE)* [1, 9]. ALE is built around the *Stella* Atari 2600 emulator. We modify OpenAI Gym and ALE to accommodate two players, and modify OpenAI Gym to allow for deploying "frozen policies" alongside policies that are still being trained. We also extend ALE by adding functionality to write to the Atari emulator's RAM, using this, for example, to give players random start positions. Our Atari framework therefore complements similar multi-agent Atari extensions (e.g., [16]); it introduces novel cooperative modes of existing games (e.g., Space Invaders), and it includes the JLE framework.

This work relates to earlier research that has studied different aspects of joint decision making in settings of two-agent collaboration, including: steering policies [6], online adaptation to the behavior of another agent (e.g., [7, 14]), repeated interactions in human-AI collaboration [4, 12], and the utility of human modeling in a collaborative game [3]. We differ in that the focus here is on studying the robustness of helper behaviors to misspecifications of partner agents.

2 CONCEPTS AND TERMINOLOGY

In explaining our results it is helpful to introduce a few different concepts and terminology. First, to train regular agents, that is players that are not explicitly designed to obtain helpful behaviors, we use ACKTR to learn to control both players, and extract and freeze single-agent policies (from a double-headed policy) at different points along a training curve. We denote these agents as S_1 through S_4, corresponding to increasing skills (S_1 is novice, and S_4 is expert level, representing training ACKTR until converged). Once trained, these agents can be evaluated in different configurations; i.e., they can be *paired with self*, e.g., $S_2 - S_2$, or paired with another type of agent, e.g., $S_2 - S_3$. We also use reward modifications to Atari games to train agents with diverse behaviors; i.e., agents that prefer

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.

	The Behavior of the Partner AI Agent					**Human Partner**	
	S_1	S_2	S_2-close	S_2-distant	S_3		
Performance with self	878	1,134	1,111	1,141	2,141	**...with** S_2	704
... with expert-skill agent (S_4)	694	963	457	711	1,826	...with S_4	545
with Helper-AI trained for different target behaviors							
$H(S_1)$	$\boxed{1,701}$	2,294	1,185	1,449	3,538		-
$H(S_2)$	1,587	$\boxed{2,434}$	1,227	1,548	3,792	...with $H(S_2)$	$\boxed{1,547}$
$H(S_2$-close)	1,254	1,836	$\boxed{1,932}$	1,405	2,733		-
$H(S_2$-distant)	1,414	2,197	1,210	$\boxed{2,375}$	3,838		-
$H(S_3)$	1,282	2,204	1,220	1,670	$\boxed{3,844}$		-
$bH(S_2)$ (a bounded helper)	1,337	2,148	1,193	1,550	3,009	...with $bH(S_2)$	1,083

Table 1: Two Player, Cooperative Space Invaders. Game score, averaged over 100 games, of pairing a partner agent (columns) with different agents (rows): whether another copy of itself, a higher-skilled agent, or a Helper-AI (both on-target and off-target). While the strongest performance comes from on-target Helper-AIs, and the worst performance comes from matching with an expert-skill agent S_4, the Helper-AI continues to provide a decisive advantage for all pairings. The Bounded-Helper-AI also provides a consistent advantage over self-pairing. The decisive performance advantage of the Helper-AIs, compared with pairing with either S_2 or S_4, holds up in transferring to this human environment.

to be close to each other or prefer to be distanced. These agents are respectively denoted by S_2-distant and S_2-close.

We train helpful agents (*Helper-AIs*) by training agents to best-respond to specific, *target behaviors*. For example, $H(S_2)$ is a Helper-AI that is trained to best-respond to S_2. Given this, $H(S_2) - S_3$ represents the configuration in which this Helper-AI is deployed along with partner S_3. Whereas Helper-AIs such as $H(S_2)$ are trained to convergence, we also train helper agents for a smaller number of episodes (*Bounded-Helper-AIs*). $bH(S_2)$, for example, results from learning to best-respond to S_2, but limiting training to the same number of episodes that are used to train S_2. In this way, $bH(S_2) - S_2$ is comparable in training effort to $S_2 - S_2$.

3 MAIN RESULTS

We summarize our main experimental results for Helper-AIs in the context of two-player, cooperative Space Invaders in Table 1.

In overview, we see that for all partner agents, the "with self" performance is worse than the performance with any of the Helper-AIs, and including the Bounded-Helper-AI and the off-target Helper-AIs, and typically substantially so. Considering human subjects, the $H(S_2)$ Helper-AI, which is trained to provide helpful behavior with a medium-skilled AI agent (S_2) also provides a substantial performance relative to pairing people with either the medium-skill AI S_2 or the expert-skill AI S_4. In some more detail, we can observe the following from these results.

Helpful behavior vs. expert behavior. Whereas pairing an agent with an expert-skill agent consistently reduces performance relative to self-pairing, there is a decisive and consistent performance improvement from pairing an AI with its on-target Helper-AI. When expressed as the percentage of score increase, the improvement from on-target Helper-AIs averages 94% across the different partner AIs, and ranges from 74% for S_2-close to 115% for S_2. For example, $H(S_2) - S_2$ scores $2,434$, $S_2 - S_2$ scores $1,134$, and $S_4 - S_2$ scores 963.

Robust helpful behavior. There is a consistent improvement in performance when pairing an AI with an off-target Helper-AI than compared to the performance from self-pairing. For example, $H(S_3) - S_2$ scores $2,204$ and $H(S_1) - S_2$ scores $2,294$, compared to just $1,134$ for $S_2 - S_2$.

Robust helpful behavior, bounded helpers. The Bounded-Helper-AI, $bH(S_2)$, provides a consistent improvement in performance for partner agents relative to self-pairing. For example, $bH(S_2) - S_2$ scores $2,148$ compared to $1,134$ for $S_2 - S_2$, and $bH(S_2) - S_2$-distant scores $1,550$ compared to $1,141$ for S_2-distant in self-pairing. This demonstrates that effective, helpful behavior that transfers to environments with off-target AIs can be learned quickly.

Robust human transfer. Helper-AIs and Bounded-Helper-AIs trained for target behavior S_2 improve performance when paired with human subjects, relative to pairing humans with medium- or high-skill non-helper AIs (e.g., $H(S_2) - Human$ is better than $S_2 - Human$). Also, pairing with the expert-level AI (S_4) actually degrades performance relative to pairing with a lower-skill AI (S_2). We ran these experiments with ten human subjects.

ACKNOWLEDGMENTS

The project or effort depicted was or is sponsored, in part, by the Defense Advanced Research Projects Agency under cooperative agreement number HR00111920029, the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Approved for public release; distribution is unlimited. This work was also supported, in part, by a SNSF Early Postdoc Mobility fellowship.

REFERENCES

[1] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.

[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540* (2016).

[3] Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the Utility of Learning about Humans for Human-AI Coordination. In *Advances in Neural Information Processing Systems*. 5175–5186.

[4] Jacob W. Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. 2018. Cooperating with Machines. *Nature Communications* 9, 1 (2018), 1–12.

[5] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. 2017. OpenAI Baselines. https://github.com/openai/baselines.

[6] Christos Dimitrakakis, David C. Parkes, Goran Radanovic, and Paul Tylkin. 2017. Multi-View Decision Processes: The Helper-AI Problem. In *Advances in Neural Information Processing Systems*. 5449–5458.

[7] Ahana Ghosh, Sebastian Tschiatschek, Hamed Mahdavi, and Adish Singla. 2020. Towards Deployment of Robust Cooperative AI Agents: An Algorithmic Framework for Learning Adaptive Policies. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 447–455.

[8] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining Cooperation in Complex Social Dilemmas using Deep Reinforcement Learning. *arXiv preprint arXiv:1707.01068* (2017).

[9] Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. 2018. Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents. *Journal of Artificial Intelligence Research* 61 (2018), 523–562.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602* (2013).

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529–533.

[12] Stefanos Nikolaidis, Swaprava Nath, Ariel D. Procaccia, and Siddhartha Srinivasa. 2017. Game-Theoretic Modeling of Human Adaptation in Human-Robot Collaboration. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 323–331.

[13] Paulo Peccin. 2015. Javatari - Online Atari 2600 Emulator. https://github.com/ppeccin/javatari.js.

[14] Goran Radanovic, Rati Devidze, David Parkes, and Adish Singla. 2019. Learning to Collaborate in Markov Decision Processes. In *International Conference on Machine Learning*. 5261–5270.

[15] 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.

[16] Justin K. Terry, Benjamin Black, and Luis Santos. 2020. Multiplayer Support for the Arcade Learning Environment. *arXiv preprint arXiv:2009.09341* (2020).

[17] Yuhuai Wu, Elman Mansimov, Shun Liao, Roger Grosse, and Jimmy Ba. 2017. Scalable Trust-Region Method for Deep Reinforcement Learning using Kronecker-Factored Approximation. In *Advances in Neural Information Processing Systems*. 5285–5294.