Title
On the Utility of Learning about Humans for Human-AI Coordination.

Authors
Carroll, Micah
Shah, Rohin
Ho, Mark K
et al.

Publication Date
2019

Peer reviewed

On the Utility of Learning about Humans for Human-AI Coordination

Part of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)

AuthorFeedback » Bibtex » Bibtex » MetaReview » Metadata » Paper » Reviews » Supplemental »

Authors

Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, Anca Dragan

Abstract

While we would like agents that can coordinate with humans, current algorithms such as self-play and population-based training create agents that can coordinate with themselves. Agents that assume their partner to be optimal or similar to them can converge to coordination protocols that fail to understand and be understood by humans. To demonstrate this, we introduce a simple environment that requires challenging coordination, based on the popular game Overcooked, and learn a simple model that mimics human play. We evaluate the performance of agents trained via self-play and population-based training. These agents perform very well when paired with themselves, but when paired with our human model, they are significantly worse than agents designed to play with the human model. An experiment with a planning algorithm yields the same conclusion, though only when the human-aware planner is given the exact human model that it is playing with. A user study with real humans shows this pattern as well, though less strongly. Qualitatively, we find that the gains come from having the agent adapt to the human's gameplay. Given this result, we suggest several approaches for designing agents that learn about humans in order to better coordinate with them. Code is available at https://github.com/HumanCompatibleAI/overcooked_ai.