Cooperative AI: A Primer

Grant Harris

Introduction

In recent years, the field of artificial intelligence (AI) has experienced tremendous growth in capabilities research to the point where AI can outperform humans at many tasks. Should this trajectory continue, it is evident that AI will soon become even more important to humanity than it already is, deployed in contexts ranging from daily life to government and corporate decision-making. Due to the imminent nature of humanity's coexistence with AI, we must ensure that the AI we develop are able to act cooperatively across a wide variety of interactions with both humans and other AI. Understanding the emerging field of cooperative AI is the first step to realizing and mitigating the risks associated with failing to develop cooperative capabilities in AI.

Although interactions involving AI and human agents and machine learning models of cooperation have been studied extensively, most recently for their applications in autonomous driving (Hsu et al. 2020, Shalev-Shwartz et al. 2016), tax policy design (Zheng et al. 2022), and multiplayer e-sports (Berner et al. 2021), the field of cooperative AI was only formalized recently (Dafoe et al. 2020). Cooperative AI research, at its broadest, focuses on "cooperative games and complex social spaces, on understanding norms and behaviours, and on social tools and infrastructure that promote cooperation" in scenarios involving one or more AI agents (Dafoe et al. 2021). Cooperation may occur on a variety of levels and with different types of agents, including humans, organizations, and other AI. The field has applications in robotics, medicine, defense, consumer technology, manufacturing, and economics – all domains that involve frequent interactions between multiple agents with related goals.

In this overview, I summarize the current state of the field of cooperative AI, discuss important related concepts and fields, and identify areas for future research. I demonstrate how understanding problems of AI cooperation and developing safe AI agents with cooperative capabilities will be useful for navigating the fast-approaching era of human-AI coexistence. Finally, I analyze the risks associated with cooperative AI and provide several examples that show how cooperative AI poses an existential threat.

Multi-Agent Learning

For dynamic environments with multiple actors engaging in complex and interrelated tasks, multi-agent systems are necessary to accurately represent the potential for interactions between agents. Multi-agent settings are useful for accurately modeling lifelike cooperation scenarios, which generally consist of autonomous actors acting within a decentralized system (Falco & Robiolo 2019). In these cases, cooperation among individual agents offers numerous benefits: increased efficiency, robustness, scalability, and reusability (Balaji & Srinivasan 2010). By sharing learned policies, sensory information, and trials from the learning process, agents can improve performance in tasks such as hunting prey (Tan 1993).

While different learning methods for AI agents exist, much of the literature about training multi-agent systems focuses on multi-agent reinforcement learning (MARL) (Buşoniu et al. 2010, Stankovic 2016, Oroojlooy & Hajinezhad 2021). In reinforcement learning, agents are rewarded for programmer-defined actions taken while in a given state; "good" actions, such as becoming more wealthy in a stock market simulation, and "bad" actions, such as moving further away from the target in a search task, are conveyed to the agents by the programmer using positive and negative rewards. Each state can be modeled as a Markov decision process, that is, agents choose from the set of all available actions with the goal of maximizing the reward (van Otterlo & Wiering 2012). MARL applies this framework to multiple agents, defining individual and/or group rewards for one or more agents. I will discuss this in greater detail in the following section.

Competitive, Cooperative, and Mixed-Motive Learning

In multi-agent settings, rewards are defined either individually for each agent or collectively for all agents. This decision determines whether the setting is competitive, cooperative, or mixed-motive, also known as general-sum (Zhang et al. 2019). Fully cooperative and fully competitive scenarios are well-defined in a programming context; however, in the real world, most scenarios fall in between these two ends of the spectrum. General-sum settings are characterized by overlapping preferences of agents, creating an environment within which agents are dependent upon each other to some degree (Schelling 1958). Any non-zero-sum game where agents have different utility functions can be characterized as general-sum. For example, a social dilemma where prey-seeking agents must choose whether to hunt alone, yielding the baseline reward upon capture, or with another agent, yielding a larger reward if both agents are nearby when the prey is captured, falls into this category (Leibo et al. 2017). In mixed-motive models, there is a tradeoff between individual and group utility that is reflected in agents' actions (Ze'evi et al. 2018).

General-sum games have recently received some consideration in the literature (Celli et al. 2019, Pérolat et al. 2017), but the cooperative actions that agents may take are often limited in these scenarios; for example, a bargaining game that only allows agents to barter for a few types of items with clearly-defined quantities and utilities for each agent (Stastny et al. 2021). While simulations like this contribute to our understanding of how agents cooperate in simple, easily-modeled environments, these environments disregard the complexity of real-world environments. More work needs to be done to model realistic non-zero-sum environments that allow for a wider variety of possible actions, and more methods and levels of cooperation.

Behavioral Game Theory

In cooperative learning, game theory plays an important role in determining how agents will act in a scenario involving different rewards given the strategies of all agents. However, negotiation between two AI differs significantly from negotiation between an AI and a human. Behavioral game theory takes into consideration human heuristics and biases that affect the decision-making process, which provides a more useful context within which to analyze human-AI interactions. Humans are by nature irrational, with incoherent preferences that often lead to actions that do not maximize expected utility (Rabin 1998). In particular, humans strongly prefer gains to losses (Kahneman & Tversky 1979). By contrast, AI act strictly to maximize expected utility and exhibit coherent preferences. As a result, humans pursue suboptimal strategies in many games, especially in complex multi-agent games with a wide range of possible actions and rewards. When a human and an AI play such games together, it follows that the AI generally outperforms the human; for example, in multi-issue negotiations between a human and an AI, the mutually agreed-upon deals benefit the AI more than the human (Dai et al. 2021).

Results from bargaining tasks also differ from the game theoretical equilibria that are standard in the literature (Nash 1950, Kalai et al. 1975), which is partially explained by the human usage of fairness, and other psychological factors that impact decision-making (Camerer 1997). A well-documented human strategy in matching-number games, where two agents are rewarded if they each guess the same number, is convergence to focal points; for example, two humans asked to choose the same number between 1 and 100 will often both guess 1 (Schelling 1958). Behavioral game theory is a subset of behavioral economics, which has made significant progress over the past few decades; however, more research is required, particularly with regard to repeated games, to formulate an accurate understanding of behavioral game theory.

Risks of Cooperative AI

While cooperative AI technology, if deployed correctly, will likely be instrumental in the progress of society, it also comes with a significant set of risks. First, combining the resources and capabilities of several agents in the pursuit of a common goal immediately demonstrates the importance of closely monitoring and controlling the learning process. If the training goes awry, and some learned behaviors are unintentionally destructive or harmful in some way, then there is a higher potential for these behaviors to spiral out of control due to the shared utility function of the cooperative agents, which prompts coordination in the planning and action stages. In this way, all risks from single-agent learning are magnified. In systems containing hundreds or thousands of agents, a harmful learned behavior would be especially catastrophic.

The issue of deploying an AI that is not sufficiently advanced enough to perform its assigned task can also pose a risk for human-AI cooperative technologies; car accidents caused by AI-assisted driving are a prime example, and are representative of the high-risk deployment scenarios for cooperative AI. Finally, adversarial attacks may cause deployed systems to perform poorly and learn suboptimal policies; obfuscating, or intentionally stalling the training or negotiation process by feeding the algorithm incorrect information, is particularly effective for cooperative AI methods that replicate human social intelligence to train agents (Fujimoto & Pederson 2021). While each of these may scale to become existential threats, the risk of combining the resources of individual agents is the most likely.

It is clear that multi-agent systems will gain prominence over time as our ability to model complex scenarios with many actors, actions, and states increases. As a result, we must expand on our current knowledge in this area to better predict and prepare for interactions between agents. Building safe forms of cooperative AI will require consensus to be reached on several issues, such as equilibrium selection, what constitutes fairness, compatibility of model beliefs, and acceptable levels of transparency (Baumann 2022). Currently, there are many approaches to cooperative AI research and little standardization or organization; greater recognition of the field and deeper understanding of partially cooperative agents would prepare humans and AI to successfully cooperate in the future.