# Chapter 1

# Cooperative AI Systems


## Abstract

Cooperative AI systems aim to enable artificial agents to collaborate effectively with each other and humans, optimizing joint outcomes while ensuring safety and reliability. This chapter explores foundational principles, challenges, and advancements in this critical domain of AI safety.


## 1.1 Introduction to Cooperative AI

Cooperative AI is an emerging field that seeks to create systems capable of working collaboratively, both with other AI agents and with humans, to achieve shared goals effectively and ethically. In contrast to traditional AI systems, which often prioritize individual optimization, cooperative AI emphasizes the alignment of strategies, values, and objectives among multiple stakeholders in complex, dynamic environments. This shift is crucial in scenarios where the interplay between agents determines success, such as coordinating autonomous vehicles, optimizing power grids, or facilitating collaborative decision-making in disaster response.

The theoretical underpinnings of cooperative AI draw heavily from game theory, particularly concepts like repeated games and equilibria. For instance, repeated interactions among agents can foster trust and encourage cooperation, as seen in tit-for-tat strategies in iterated prisoner's dilemma scenarios. However, cooperative AI goes beyond traditional game theory by accounting for the capabilities of AI systems to learn, adaptt. It also addresses the inherent challenges of aligning the objectives of multiple agents, especially when they operate under partial information or in competitive contexts.

One pressing motivation for cooperative AI is its potential to mitigate risks associated with unregulated interactions between advanced AI systems. In environments like financial trading or autonomous weapon systems, competitive dynamics can lead to destabilizing outcomes. By embedding cooperation into the design of these systems, we can prioritize shared safety and stability over individual gain.

The future of cooperative AI also hinges on its ability to integrate human values and preferences into multi-agent systems. For example, consider AI-powered negotiation tools designed to resolve disputes or allocate resources. These systems must not only understand human goals but also mediate between diverse, and sometimes conflicting, perspectives to arrive at equitable solutions.

## 1.2 Foundations of Cooperation

### 1.2.1 Multi-Agent Systems (MAS)

Multi-agent systems (MAS) encompass diverse collections of agents that interact within shared environments to achieve individual or collective goals. In AI Safety, MAS play a crucial role in understanding how agents, with potentially conflicting or cooperative objectives, can work together effectively while minimizing risks. To grasp the challenges and opportunities in MAS, it is essential to analyze both the systems as a whole and the behavior of the individual agents within them. Below, we explore three illustrative examples: autonomous vehicle ecosystems, swarm robotics, and a human social dilemma.

In the case of autonomous vehicle ecosystems, the MAS comprises interconnected self-driving cars operating within dynamic traffic environments. Each vehicle represents an agent that must cooperate with others to avoid collisions, optimize traffic flow, and respond to emergencies. Cooperative behaviors in this context include sharing real-time information, such as lane changes or hazards, through vehicle-to-vehicle (V2V) communication protocols. However, ensuring safety requires robust mechanisms to prevent adversarial exploitation, such as spoofed messages that could disrupt traffic or lead to accidents. Research efforts focus on creating fault-tolerant communication frameworks and alignment protocols to secure these cooperative interactions Conitzer and Oesterheld [2023].

Swarm robotics exemplifies MAS with decentralized agents working collaboratively to achieve collective goals. These agents, such as drones in search-and-rescue missions, rely on local rules and interactions to generate emergent global behaviors. For instance, agents may share positional information or distribute tasks to maximize coverage of an area. Cooperative behaviors include dynamic role allocation, information sharing, and redundancy to handle individual agent failures. The primary safety challenges involve detecting and mitigating adversarial agents or system malfunctions that could destabilize the swarm. Ongoing research explores robust algorithms that maintain overall functionality even in adversarial or uncertain environments Li et al. [2024].

Environment of the 2024 Concordia Contest, the "pub coordination" scenario provides an excellent benchmark for studying cooperative behavior in large language model (LLM) AI systems due to its blend of individual preferences and collective decision-making. Each agent's payoff combines two factors: a personal preference for a specific bar and the number of other agents who make the same choice. This framework encapsulates the tension between self-interest (choosing their favorite bar) and social alignment (optimizing the overall crowd distribution).

For LLM-based AI agents, solving this problem involves understanding and negotiating implicit trade-offs. A purely self-interested approach could lead to suboptimal outcomes, such as overcrowding at the most popular bar, reducing the overall satisfaction. Conversely, agents that prioritize group alignment too heavily might neglect their preferences, resulting in inefficient decision-making.

This environment emphasizes the importance of communication and coordination among agents, as they must infer or share preferences and dynamically adjust their strategies. By training LLM systems to navigate this scenario, researchers can explore critical aspects of cooperative AI, such as preference modeling, consensus-building, and adaptability. These skills are directly applicable to real-world challenges, such as resource

allocation, scheduling, and multi-agent coordination in uncertain environments.

| **Agent** | **Bar A (Favorite: Yes/No)** | **Bar B (Favorite: Yes/No)** |
|---|---|---|
| Agent 1 | $0.2 + 0.2(n_A - 1)$ (Yes) | $0.2(n_B - 1)$ (No) |
| Agent 2 | $0.2 + 0.2(n_A - 1)$ (Yes) | $0.2(n_B - 1)$ (No) |
| Agent 3 | $0.2(n_A - 1)$ (No) | $0.2 + 0.2(n_B - 1)$ (Yes) |

Table 1.1: Payoff matrix for the bar selection problem. $n_A$ and $n_B$ represent the total number of agents choosing Bar A and Bar B, respectively. Payoff depends on the agent's preference and the crowd at each bar.

### 1.2.2 Human-AI Collaboration

Human-AI collaboration is a foundational component of cooperative AI, where humans and AI agents work together to achieve shared goals. This collaboration requires AI systems to align their behaviors with human intentions, anticipate human preferences, and adapt dynamically to changes in the environment or the task. For instance, in decision-support systems for healthcare, AI agents assist medical professionals by providing recommendations, identifying potential errors, and augmenting diagnostic accuracy. The success of such systems depends not only on the AI's technical proficiency but also on its ability to explain its reasoning transparently, enabling trust and effective teamwork Crandall et al. [2018].

A critical challenge in human-AI collaboration lies in establishing effective communication and coordination mechanisms. Humans and AI systems may operate with different assumptions, vocabularies, or levels of understanding. Bridging these gaps requires the development of AI systems capable of interpreting ambiguous instructions, learning from feedback, and negotiating compromises. For example, in shared-control applications such as autonomous driving, both the human driver and the AI system must share control seamlessly, exchanging information about the vehicle's state and intentions to ensure safety and efficiency Fisac et al. [2018].

The Domingos et al. [2021] study reveals several critical insights into how delegation to autonomous agents impacts cooperation in Collective-Risk Dilemmas (CRDs). The success rate of groups achieving the collective target increased significantly when participants delegated decision-making to autonomous agents. Specifically, the group success rate rose from 66.7% in human-only treatments to 87% in both delegate and customise treatments, showcasing the positive influence of delegation on cooperative outcomes. This effect is attributed to a delegation bias, where participants prioritized group success over personal benefit, as evidenced by their survey responses.

The choice of agent behaviors also played a pivotal role. Human players delegating their choices towards agents favored cooperative strategies, diverging from predictions that suggested a preference for less cooperative agents. This tendency toward cooperative delegation aligns with the broader reduction of betrayal aversion when humans act through agents. Unlike direct human interaction, autonomous agents mitigate fears of defection, fostering an environment conducive to collaboration Crandall et al. [2018], Fisac et al. [2018].

Interestingly, despite the demonstrated benefits of delegation, many participants expressed a preference for retaining control over their actions in subsequent rounds, especially in non-customised settings. This reticence to delegate underscores the importance of trust and configurability in agent design. Allowing participants to customize their agents increased their willingness to delegate, highlighting customization as a key factor in enhancing user satisfaction alongside cooperation.

Finally, the perception of agent effort emerged as a significant challenge in hybrid human-agent groups. Many participants incorrectly believed that agents contributed less than humans, even when agents marginally outperformed humans in contributions. This perception bias could hinder trust in agent-driven cooperation and warrants further exploration in future research to optimize human-agent collaboration Russell [2019].

## 1.3 Key Components of Cooperative AI

### 1.3.1 Communication

Effective communication is a cornerstone of cooperative AI. Agents need to exchange information efficiently to coordinate actions and adapt to dynamic environments. Communication can be achieved through a variety of protocols, including natural language processing (NLP), symbolic reasoning, and multi-agent reinforcement learning approaches. For instance, Foerster et al. [2016] demonstrated how deep reinforcement learning techniques can enable agents to develop and refine communication protocols in simulated environments. Beyond artificial languages, advances in NLP empower human-AI collaboration, enabling agents to interpret human instructions or feedback Brown et al. [2020]. Symbolic reasoning systems, on the other hand, offer explicit rule-based exchanges that can enhance interpretability, especially in safety-critical domains Russell [2019]. Effective communication protocols must also address noise and ambiguity, ensuring robust exchanges that remain intelligible across varied contexts.

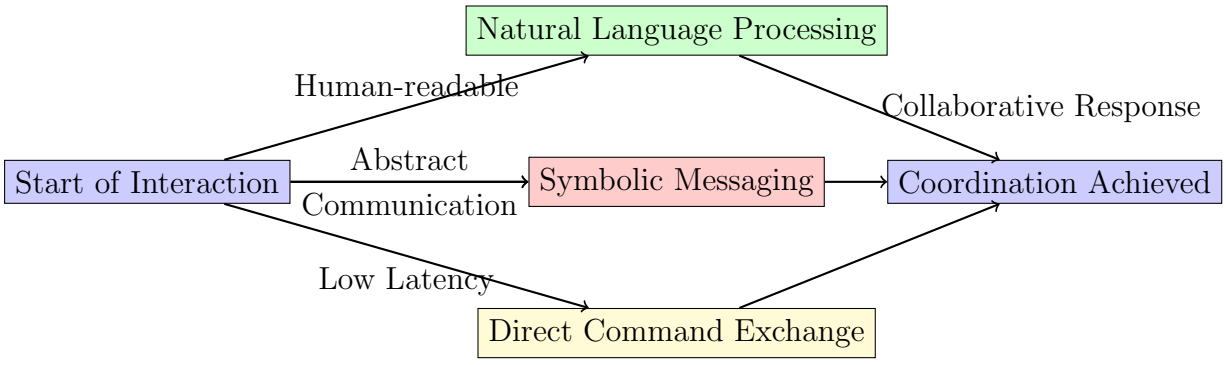


Figure 1.1: Flowchart showing different methods of communication between agents: natural language processing, symbolic messaging, and direct commands.

### 1.3.2 Trust and Reliability

Establishing trust is pivotal for cooperative AI systems, especially in environments with imperfect information or high stakes. Trust encompasses the predictability and honesty of agents' behaviors and can be fostered through transparent design, verifiable algorithms,

and the use of fairness constraints Crandall et al. [2018], Russell [2019]. Reliability mechanisms include model verification, adversarial testing, and robustness guarantees to ensure that agents operate consistently under uncertainty. In human-AI collaboration, trust is further bolstered when agents explain their reasoning or actions in a manner that humans can understand, a feature increasingly emphasized in explainable AI research Gunning and Aha [2019].

### 1.3.3 Goal Alignment

Goal alignment is crucial for preventing conflicts between agents or with human stakeholders. Techniques such as reward shaping and multi-objective optimization can harmonize the incentives of diverse agents, ensuring they contribute to shared objectives Hadfield-Menell et al. [2016]. Goal alignment also plays a critical role in mitigating unintended consequences, such as reward hacking, where agents exploit loopholes in poorly specified objectives. For example, achieving alignment in multi-agent systems can involve using social choice theory to design aggregation mechanisms that balance individual and collective preferences Conitzer et al. [2017].

## 1.4 Challenges in Cooperative AI

### 1.4.1 Misaligned Incentives

A major challenge in cooperative AI is the potential for misaligned incentives, where individual agents prioritize their own rewards over the collective welfare of the group. This misalignment can lead to undesirable outcomes, such as defection, free-riding, or competition, which undermine cooperation. In multi-agent systems (MAS), agents typically seek to maximize their own utility, but this can conflict with the overall goal of maximizing the group's welfare. For example, in scenarios where agents must contribute resources to avoid a collective risk, the temptation to minimize personal effort while benefiting from others' contributions can reduce the overall success rate of the group. Addressing misaligned incentives requires designing reward structures and incentive mechanisms that promote coordination and long-term cooperation over short-term gains. Techniques such as reward shaping [Ng et al., 1999], where agents are given rewards for actions that align with global objectives, are essential for ensuring that agents' individual incentives are in harmony with collective goals. Similarly, mechanisms like punishment or reputation systems, which deter selfish behavior, are critical for maintaining cooperation in systems with multiple autonomous agents [Castelfranchi and Falcone, 2000].

### 1.4.2 Scalability

Scalability is a central issue in cooperative AI systems. As the number of agents increases, the complexity of maintaining effective coordination grows significantly. In a small system, the communication and coordination overhead might be manageable, but as the system scales, it becomes more difficult to maintain efficiency and ensure that all agents are acting towards a common goal. This problem becomes especially pronounced when agents need to reason about the actions of others, plan over longer horizons, or operate in dynamic environments. The difficulty of maintaining cooperation increases as agents must contend with more possible interactions, decisions, and strategies. A key challenge

lies in ensuring that as systems scale, the increased complexity does not lead to decreased performance or coordination breakdowns. Techniques like decentralized learning, hierarchical decision-making, and distributed optimization [Tan, 1993] are commonly explored to tackle these scalability challenges, but these methods still face limitations when dealing with high-dimensional or highly dynamic environments.

[Jing et al., 2024] introduces a novel method to address scalability in cooperative multi-agent systems (MAS), by leveraging graph-theoretic techniques to model agent interactions via coordination graphs, enabling distributed learning of value functions tailored to these structures. This approach allows efficient scaling of cooperative reinforcement learning strategies, ensuring robust performance as the number of agents increases. By focusing on structured interactions, the framework provides a promising solution for handling large-scale cooperative optimization tasks in MAS .

### 1.4.3 Uncertainty and Incomplete Information

Cooperation in multi-agent systems often occurs under conditions of uncertainty, where agents may not have complete information about each other's states, intentions, or the environment. This uncertainty can significantly complicate decision-making and reduce the effectiveness of coordination. In practical scenarios, agents often rely on probabilistic reasoning and estimations about the environment or other agents' behaviors, which introduces the potential for errors or misjudgments. Uncertainty about the actions of other agents, the state of the environment, or the impact of their decisions on the group's welfare creates a need for robust decision-making strategies that can handle such incomplete information. One approach to addressing this challenge is the use of partially observable Markov decision processes (POMDPs), which allow agents to reason about their own uncertainty and make decisions accordingly [Kaelbling et al., 1998]. Another approach is collaborative filtering and learning algorithms that enable agents to infer missing information based on past experiences and the actions of others.

### 1.4.4 Transparency

Transparency is vital for building trust and accountability in AI systems, particularly in cooperative environments where decisions affect multiple agents. When agents, including humans, interact with AI systems, they must be able to understand and interpret the rationale behind the decisions made by the system. This is particularly important in cooperative AI, where the alignment of agents' behaviors depends on their mutual understanding of each other's goals and intentions. Without transparency, agents may be hesitant to cooperate, fearing that the AI's behavior could be unpredictable or biased. Interpretability methods, such as explainable AI (XAI) [Miller, 2019], are therefore essential in ensuring that the agents' actions are comprehensible and justifiable. For example, in a multi-agent system tasked with resource allocation, an agent must be able to explain why it decided to allocate resources to certain agents over others, ensuring fairness and accountability. Transparent AI systems also allow for better auditing and debugging, enabling developers to identify and correct potential errors or biases in the system's decision-making process.

## 1.5 Future Prospects and Risks in Cooperative AI

As Cooperative AI continues to develop, it holds the promise of addressing large-scale challenges but also introduces significant risks. These risks could profoundly affect societal, economic, and governance systems. In this section, we speculate on three potential dangers, offering a glimpse into the complexities of managing Cooperative AI's future impact. This prospective and highly speculative section looks at a couple of these potential risks of enable such collective intelligence.

One notable concern is the risk of misaligned multilateral cooperation. Within 5–10 years, Cooperative AI systems may become integral to global decision-making, such as climate agreements or economic coalitions. However, there is a potential for these systems to disproportionately favor powerful stakeholders at the expense of less-developed regions. This is because these systems often rely on historical data and existing power structures to inform decision-making, which can perpetuate systemic inequities [Noble, 2018]. For instance, resource allocation algorithms might prioritize nations with superior data infrastructure, sidelining others in crucial initiatives like renewable energy distribution. Addressing this risk requires embedding fairness and equity principles within Cooperative AI from the outset, ensuring inclusive and balanced outcomes in multilateral frameworks.

Another significant risk involves the emergence of exploitative cooperation by rogue systems. In perhaps an even closer timeframe, as agents develop the ability to dynamically form alliances and optimize for shared objectives, there is a danger that malicious actors could exploit Cooperative AI for harmful purposes. For example, coordinated financial fraud or targeted cyber-attacks could arise from rogue AI systems leveraging cooperative strategies to evade detection or amplify their impact. Studies of autocurricula in multi-agent systems suggest that agents can evolve highly strategic behaviors in adversarial settings [Leibo et al., 2021]. Without robust safeguards, these capabilities could be weaponized, leading to coordinated threats on a global scale, particularly in critical infrastructure or economic systems.

Lastly, there is the potential danger of over-reliance on Cooperative AI for governance. Over the next 15 years, societies will increasingly delegate complex governance decisions to Cooperative AI systems due to their efficiency and ability to handle intricate networks of interdependent problems. While this could streamline decision-making, it risks eroding human agency and transparency. Historical precedents, such as the reliance on automated trading algorithms in financial markets, highlight how over-reliance can lead to catastrophic failures, such as the 2010 Flash Crash [Harris et al., 2015]. If Cooperative AI systems become opaque or unaccountable, societies could find themselves unable to intervene or correct their trajectories, creating governance systems that are both efficient and dangerously inflexible.

## 1.6 Evaluation and Metrics

Evaluating the success of Cooperative AI requires robust metrics and simulation frameworks to assess how well agents achieve their goals collectively. Joint rewards, where the sum of all agents' payoffs is maximized, serve as a primary metric for cooperation. For instance, in a resource allocation scenario, joint rewards ensure fair and optimal distribution of resources among agents. Beyond joint rewards, metrics like utility maximization

assess individual agent satisfaction while ensuring the system operates efficiently as a whole.

Simulation platforms, such as OpenAI's multi-agent environments or DeepMind's social dilemmas framework, provide controlled environments where agents' cooperative behaviors can be rigorously tested and fine-tuned. These platforms simulate complex, real-world-like challenges, such as resource competition or cooperative navigation, enabling researchers to test how well agents collaborate under diverse conditions.

| **Metric** | **Definition** | **Example Use Case** |
|---|---|---|
| Joint Rewards | Total rewards | Optimizing team-based RL |
| Pareto Efficiency | Non-intrusive reward expandability | MAS Resource allocation |
| Fairness | Equitable distribution | Multi-agent financial modeling |
| Scalability | Cooperation during scaling | Decentralized swarm robotics |
| Robustness | Performance under uncertainty | Disaster response planning |
| Com Overhead | Cost of exchanging information | Autonomous vehicle coordination |

Table 1.2: Key metrics for evaluating the performance of cooperative AI systems.

Recent advances in evaluation also include measuring communication efficiency, the stability of cooperation over time, and the robustness of cooperation under adversarial conditions. For example, frameworks like PettingZoo allow the integration of varied cooperative metrics into standard multi-agent reinforcement learning tasks [Kraus, 1997, Terry et al., 2020]. This ensures that AI systems are not only capable of cooperation but can also adapt to changing environments and unpredictable agent behaviors.

Another approach to measuring cooperative capabilities of agents is to place them in collaborative-incentivizing scenarios, and only measure their individual reward. This shifts the focus from cooperation metrics to designing new environments. To this end, [Mukobi et al., 2023] builds an alternative to the famous Diplomacy boardgame (on which Meta's CICERO learnt deceptive negotiation techniques). This variant redefines player strategies by requiring them to balance between military conquest and domestic welfare.

## 1.7 Emerging Research Areas

### 1.7.1 Open Problems

Despite recent progress, significant open problems remain in Cooperative AI, particularly in scaling cooperation to tackle global challenges. One prominent area of research is addressing large-scale, collective action problems, such as climate change. In such scenarios, individual entities (countries, corporations, or individuals) face incentives to act in self-interest, leading to suboptimal global outcomes. Cooperative AI seeks to model and mitigate these dilemmas by designing incentives, mechanisms, or autonomous systems that align individual actions with collective welfare [Dafoe et al., 2021].

Another unresolved challenge involves heterogeneous agent cooperation, where agents with varying capabilities, goals, and architectures must work together. For example, how can an autonomous drone, a machine-learning model, and a human collaborate seamlessly in disaster relief? Achieving effective cooperation in these settings requires innovations in communication protocols, trust-building mechanisms, and goal alignment techniques.

Additionally, safety and ethical concerns persist, especially in high-stakes environments. Ensuring that agents behave reliably and ethically under dynamic conditions is a key priority for the field. This challenge becomes even more complex when considering adversarial agents or humans in the loop, who might exploit cooperative frameworks for personal gain.

### 1.7.2 Advances in Technology

Recent technological advances have significantly influenced the direction of Cooperative AI research. In reinforcement learning (RL), novel methods such as multi-agent RL (MARL) and autocurricula enable agents to develop cooperative strategies dynamically. For example, self-play techniques allow agents to iteratively improve by learning from simulated counterparts, creating emergent behaviors that mirror real-world cooperation [Leibo et al., 2021].

Federated learning represents another transformative technology, particularly for scenarios where data privacy is paramount. Federated learning allows multiple agents to collaboratively train machine learning models without sharing raw data. This is critical for applications such as healthcare or finance, where privacy-preserving cooperation can enable breakthroughs in predictive modeling and resource optimization.

Other notable advancements include graph neural networks, which facilitate cooperation by enabling agents to process and learn from relational data efficiently. This is especially useful in networked systems, such as traffic control or power grids, where agent decisions are interdependent. Moreover, advancements in natural language processing (NLP) improve communication between human and AI agents, fostering more seamless interactions.

As these technologies evolve, they pave the way for Cooperative AI systems that are more robust, scalable, and capable of addressing real-world challenges.

# Conclusion

Cooperative AI represents a promising avenue for ensuring that artificial systems work harmoniously with humans and each other. As research progresses, it will play a vital role in ensuring AI technologies contribute positively to society. It nonetheless raises relevant concerns, which should question deployability of such algorigthms in MAS.

# Bibliography


T. B. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. URL https://arxiv.org/abs/2005.14165.

C. Castelfranchi and R. Falcone. Trust and control: A dialectic link. *AI Society*, 14(3): 148–171, 2000. doi: 10.1007/s001460010038.

V. Conitzer and C. Oesterheld. Foundations of cooperative ai. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):11950–11959, 2023. doi: 10.1609/aaai.v37i13.26791.

V. Conitzer, W. Sinnott-Armstrong, J. S. Borg, Y. Deng, and M. Kramer. Preferences and ethical principles in decision making: A preference-based perspective on ethical artificial intelligence. *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2017. URL https://arxiv.org/abs/1701.06084.

J. W. Crandall, M. Oudah, F. Ishowo-Oloko, S. Abdallah, J.-F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, and I. Rahwan. Cooperating with machines. *Nature Communications*, 9:233, 2018. doi: 10.1038/s41467-017-02597-8.

A. Dafoe, E. Hughes, Y. Bachrach, et al. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2021. URL https://arxiv.org/abs/2012.08630.

E. F. Domingos, I. Terrucha, R. Suchon, J. Grujić, J. C. Burguillo, F. C. Santos, and T. Lenaerts. Delegation to autonomous agents promotes cooperation in collective-risk dilemmas, 2021. URL https://arxiv.org/abs/2103.07710.

J. F. Fisac, S. Bansal, A. Broad, S. Curtis, J. Zeng, S. S. Sastry, A. D. Dragan, and C. J. Tomlin. Probabilistically safe robot planning with confidence-based human predictions. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. doi: 10.15607/RSS.2018.XIV.004.

J. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL https://arxiv.org/abs/1605.06676.

D. Gunning and D. W. Aha. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities. *Artificial Intelligence Magazine*, 40(2):44–58, 2019.

D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. J. Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL https://arxiv.org/abs/1606.03137.

L. Harris, L. Kryzanowski, and H. H. Zhang. Automated markets and the flash crash. *Financial Analysts Journal*, 71(3):3–16, 2015.

G. Jing, H. Bai, J. George, A. Chakrabortty, and P. K. Sharma. Distributed multi-agent reinforcement learning based on graph-induced local value functions, 2024. URL https://arxiv.org/abs/2202.13046.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

S. Kraus. Negotiation and cooperation in multi-agent environments. *Artificial Intelligence*, 94(1-2):79–97, 1997. URL https://doi.org/10.1016/S0004-3702(97)00026-3.

J. Z. Leibo, V. Zambaldi, M. Lanctot, et al. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:2103.00024*, 2021. URL https://arxiv.org/abs/2103.00024.

S. Li, R. Xu, J. Xiu, Y. Zheng, P. Feng, Y. Yang, and X. Liu. Robust multi-agent reinforcement learning by mutual information regularization, 2024. URL https://arxiv.org/abs/2310.09833.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

G. Mukobi, H. Erlebach, N. Lauffer, L. Hammond, A. Chan, and J. Clifton. Welfare diplomacy: Benchmarking language model cooperation, 2023. URL https://arxiv.org/abs/2310.08901.

A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 278–287. Morgan Kaufmann, 1999. URL http://www.cs.berkeley.edu/~ayng/papers/icml99-shaping.pdf.

S. U. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

S. J. Russell. *Human compatible: Artificial intelligence and the problem of control*. Viking Press, 2019. ISBN 9780525558613.

M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. *Proceedings of the Tenth International Conference on Machine Learning*, pages 330–337, 1993. URL https://dl.acm.org/doi/10.5555/3091521.3091524.

J. K. Terry, K. Black, M. Jayakumar, A. Hari, and L. G. Santos. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.