Cooperative AI Foundation
Technical Report #1
February 2025

Multi-Agent Risks from Advanced AI

Lewis Hammond
Alan Chan
Jesse Clifton
Jason Hoelscher-Obermaier
Akbir Khan
Euan McLean
Chandler Smith
Wolfram Barfuss
Jakob Foerster
Tomáš Gavenčiak
The Anh Han
Edward Hughes
Vojtěch Kovařík
Jan Kulveit
Joel Z. Leibo
Caspar Oesterheld
Christian Schroeder de Witt
Nisarg Shah
Michael Wellman
Paolo Bova
Theodor Cimpeanu
Carson Ezell
Quentin Feuillade-Montixi
Matija Franklin
Esben Kran
Igor Krawczuk
Max Lamparth
Niklas Lauffer
Alexander Meinke
Sumeet Motwani
Anka Reuel
Vincent Conitzer
Michael Dennis
Iason Gabriel
Adam Gleave
Gillian Hadfield
Nika Haghtalab
Atoosa Kasirzadeh
Sébastien Krier
Kate Larson
Joel Lehman
David C. Parkes
Georgios Piliouras
Iyad Rahwan

Multi-Agent Risks from Advanced AI

Lewis Hammond[1,2,*]

Alan Chan[3,4], Jesse Clifton[5,1], Jason Hoelscher-Obermaier[6], Akbir Khan[7,8,(1)], Euan McLean[†], Chandler Smith[1]

Wolfram Barfuss[9], Jakob Foerster[2,10], Tomáš Gavenčiak[11], The Anh Han[12], Edward Hughes[13], Vojtěch Kovařík[14,(15)], Jan Kulveit[11], Joel Z. Leibo[13], Caspar Oesterheld[15], Christian Schroeder de Witt[2], Nisarg Shah[16], Michael Wellman[17]

Paolo Bova[12], Theodor Cimpeanu[18,(19)] Carson Ezell[20], Quentin Feuillade-Montixi[21,(†)], Matija Franklin[8], Esben Kran[6], Igor Krawczuk[†,(22)], Max Lamparth[23], Niklas Lauffer[24], Alexander Meinke[25,(†)], Sumeet Motwani[2,(24)], Anka Reuel[23,20]

Vincent Conitzer[15], Michael Dennis[13], Iason Gabriel[13], Adam Gleave[26], Gillian Hadfield[27], Nika Haghtalab[24], Atoosa Kasirzadeh[15], Sébastien Krier[13], Kate Larson[28,13], Joel Lehman[†], David C. Parkes[20], Georgios Piliouras[13,29], Iyad Rahwan[30]

[1]Cooperative AI Foundation
[2]University of Oxford
[3]Mila
[4]Centre for the Governance of AI
[5]Center on Long-Term Risk
[6]Apart Research
[7]Anthropic
[8]University College London
[9]University of Bonn
[10]Meta AI
[11]Charles University
[12]Teesside University
[13]Google DeepMind
[14]Czech Technical University
[15]Carnegie Mellon University
[16]University of Toronto
[17]University of Michigan
[18]University of Stirling
[19]University of St Andrews
[20]Harvard University
[21]PRISM AI
[22]École Polytechnique Fédérale de Lausanne
[23]Stanford University
[24]University of California, Berkeley
[25]Apollo Research
[26]FAR.AI
[27]Johns Hopkins University
[28]University of Waterloo
[29]Singapore University of Technology and Design
[30]Max Planck Institute for Human Development
[†]Independent

Abstract

The rapid development of advanced AI agents and the imminent deployment of many instances of these agents will give rise to multi-agent systems of unprecedented complexity. These systems pose novel and under-explored risks. In this report, we provide a structured taxonomy of these risks by identifying three key failure modes (miscoordination, conflict, and collusion) based on agents' incentives, as well as seven key risk factors (information asymmetries, network effects, selection pressures, destabilising dynamics, commitment problems, emergent agency, and multi-agent security) that can underpin them. We highlight several important instances of each risk, as well as promising directions to help mitigate them. By anchoring our analysis in a range of real-world examples and experimental evidence, we illustrate the distinct challenges posed by multi-agent systems and their implications for the safety, governance, and ethics of advanced AI.

*Correspondence to lewis.hammond@cooperativeai.org. Suggested citation: "Hammond et al. (2025). *Multi-Agent Risks from Advanced AI*. Cooperative AI Foundation, Technical Report #1." Author clusters are ordered by approximate magnitude of contribution and represent the lead author, organisers, major contributors, minor contributors, and advisors, respectively. Within clusters, authors are listed alphabetically. Full details of author roles are available in Appendix A. Affiliations in parentheses indicate that the author's work on this report was primarily completed while under that affiliation. Due to the length of the author list, authorship does not entail endorsement of all claims in the report, nor does inclusion entail an endorsement on the part of any individual's organisation. In particular, contributions to this report reflect the views of the respective contributors and not necessarily the views of the Cooperative AI Foundation, its trustees, or funders.

Contents

Executive Summary

The proliferation of increasingly advanced AI not only promises widespread benefits, but also presents new risks (Bengio et al., 2024; Chan et al., 2023). **Today, AI systems are beginning to autonomously interact with one another and adapt their behaviour accordingly, forming *multi-agent systems*.** This change is due to the widespread adoption of sophisticated models that can interact via a range of modalities (including text, images, and audio), and the competitive advantages conferred by autonomous, adaptive agents (Anthropic, 2024a; Google DeepMind, 2024; OpenAI, 2025).

While still relatively rare, groups of advanced AI agents are already responsible for tasks that range from trading million-dollar assets (AmplifyETFs, 2025; Ferreira et al., 2021; Sun et al., 2023a) to recommending actions to commanders in battle (Black et al., 2024; Manson, 2024; Palantir, 2025). **In the near future, applications will include not only economic and military domains, but are likely to extend to energy management, transport networks, and other critical infrastructure** (Camacho et al., 2024; Mayorkas, 2024). Large populations of AI agents will also feature in more familiar social settings as intelligent personal assistants or representatives, capable of being delegated increasingly complex and important tasks.

While bringing new opportunities for scalable automation and more diffuse benefits to society, **these advanced, multi-agent systems present novel risks that are distinct from those posed by *single agents* or *less advanced* technologies**, and which have been systematically underappreciated and understudied. This lack of attention is partly because present-day multi-agent systems are rare (and those that do exist are often highly controlled, such as in automated warehouses), but also because even single agents present many unsolved problems (Amodei et al., 2016; Anwar et al., 2024; Hendrycks et al., 2021). Given the current rate of progress and adoption, however, **we urgently need to evaluate (and prepare to mitigate) multi-agent risks from advanced AI**. More concretely, we provide recommendations throughout the report that can largely be classified as follows.

- **Evaluation**: Today's AI systems are developed and tested in isolation, despite the fact that they will soon interact with each other. In order to understand how likely and severe multi-agent risks are, we need new methods of detecting how and when they might arise, such as: evaluating the cooperative capabilities, biases, and vulnerabilities of models; testing for new or improved dangerous capabilities in multi-agent settings (such as manipulation, collusion, or overriding safeguards); more open-ended simulations to study dynamics, selection pressures, and emergent behaviours; and studies of how well these tests and simulations match real-world deployments.
- **Mitigation**: Evaluation is only the first step towards mitigating multi-agent risks, which will require new technical advances. While our understanding of these risks is still growing, there are promising directions that we can begin to explore now, such as: scaling peer incentivisation methods to state-of-the-art models; developing secure protocols for trusted agent interactions; leveraging information design and the potential transparency of AI agents; and stabilising dynamic multi-agent networks and ensuring they are robust to the presence of adversaries.
- **Collaboration**: Multi-agent risks inherently involve many different actors and stakeholders, often in complex, dynamic environments. Greater progress can be made on these interdisciplinary problems by leveraging insights from other fields, such as: better understanding the causes of undesirable outcomes in complex adaptive systems and evolutionary settings; determining the moral responsibilities and legal liabilities for harms not caused by any single AI system; drawing lessons from existing efforts to regulate multi-agent systems in high-stakes contexts, such as financial markets; and determining the security vulnerabilities and affordances of multi-agent systems.

To support these recommendations, **we introduce a taxonomy of AI risks that are new, much more challenging, or qualitatively different in the multi-agent setting, together with a preliminary assessment of what can be done to mitigate them**. We identify three high-level *failure modes*, which depend on the nature of the agents' objectives and the intended behaviour of the system: miscoordination, conflict, and collusion. We then describe seven key *risk factors* that can lead to these failures: information asymmetries, network effects, selection pressures, destabilising dynamics, commitment and trust, emergent agency, and multi-agent security. For each problem we provide a *definition*, key *instances* of how and where it can arise, illustrative *case studies*, and promising *directions* for future work. We conclude by discussing the *implications* for existing work in AI safety, AI governance, and AI ethics.

1 Introduction

The proliferation of increasingly advanced AI not only promises widespread benefits, but also presents new risks (Bengio et al., 2024; Chan et al., 2023). In the future, AI systems will commonly interact and adapt in response to one another, forming *multi-agent systems*.[1] This trend will be driven by several factors. First, recent technical progress and publicity will continue to drive adoption, including in high-stakes areas such as financial trading (AmplifyETFs, 2025; Ferreira et al., 2021; Sun et al., 2023a) and military strategy (Black et al., 2024; Manson, 2024; Palantir, 2025). Second, AI systems that can act autonomously and adapt while deployed as *agents* will have competitive advantages compared to non-adaptive systems or those with humans in the loop. Third, the more widely such agents are deployed, the more they will come to interact with one another.

The emergence of these advanced multi-agent systems presents a number of risks which have thus far been systematically underappreciated and understudied. In part, this lack of attention is because the deployment of such systems is currently rare, or constrained to highly controlled settings (such as automated warehouses) that do not suffer from the most severe risks. In part, it is because even the simpler problem of ensuring the safe and ethical behaviour of a *single* advanced AI system is far from solved (Amodei et al., 2016; Anwar et al., 2024; Hendrycks et al., 2021), and multi-agent settings are strictly more complex. Indeed, many multi-agent risks are inherently sociotechnical and require attention from many stakeholders and researchers across many disciplines (Curtis et al., 2024; Lazar & Nelson, 2023).

Importantly, these risks are distinct from those posed by *single agents* or *less advanced* technologies, and will not necessarily be addressed by efforts to mitigate the latter. For example: the alignment of AI agents with different actors is insufficient to prevent conflict if those actors have diverging interests (Critch & Krueger, 2020; Dafoe et al., 2020; Jagadeesan et al., 2023a; Manheim, 2019; Sourbut et al., 2024); errors that may be acceptable in isolation could compound in complex, dynamic networks of agents (Buldyrev et al., 2010; Kirilenko et al., 2017; Lee & Tiwari, 2024; Maas, 2018; Sanders et al., 2018); and groups of agents could combine or collude to develop dangerous capabilities or goals that cannot be ascribed to any individual (Calvano et al., 2020; Drexler, 2022; Jones et al., 2024; Mogul, 2006; Motwani et al., 2024). Advanced AI also introduces phenomena that differ fundamentally from previous generations of AI or other technologies, requiring new approaches to mitigating these risks (Bengio et al., 2024).

With the current rate of progress, we therefore urgently need to evaluate (and prepare to mitigate) multi-agent risks from advanced AI. In this report we take a first step in this direction by providing a taxonomy of risks that either: emerge, are much more challenging, or are qualitatively different in the multi-agent setting (see Table 1). We identify three key high-level **failure modes** (Section 2), and seven key **risk factors** that can lead to these failures (Section 3), before discussing the **implications** for AI safety, AI governance, and AI ethics (Section 4). Throughout the report we illustrate these risks with concrete examples, either from real-world events, previous research, or novel experiments (see Table 3).

1.1 Overview

We begin by identifying different *failure modes* in multi-agent systems based on the nature of the agents' goals and the intended behaviour of the system. In most multi-agent systems, we are interested in AI agents working together to achieve their respective goals or the goals of those who deployed them. In this case, we categorise failures into **miscoordination** (Section 2.1), where agents fail to cooperate despite having the *same goal*, and **conflict** (Section 2.2), where agents with *different goals* fail to cooperate. A third and final kind of failure – **collusion** (Section 2.3) – can arise in competitive settings where we do *not* want agents cooperating (such as markets).

We next introduce a number of *risk factors* by which these failure modes can arise, and which are largely independent of the agents' precise incentives.[2] For example, information asymmetries could lead to miscoordination between agents with the same goal, or to conflict among agents with competing goals. These factors are not specific to AI systems, but the differences between AI systems and other kinds of intelligent agents (such as humans or corporations) leads to different risk instances and potential solutions. Finally, note that the following factors are not necessarily exhaustive or mutually exclusive.

[1] A fundamental fact about (software-based) AI systems is that they can be easily duplicated. Thus, the vast training costs involved in producing state-of-the-art systems can be amortized over millions of instances. In this sense, if nothing else, the concept of multi-agent systems is core to transformative AI.

[2] Indeed, there are potential risks from multi-agent systems in which it is not the agents' objectives that are the critical feature, but their general incompetencies or vulnerabilities.

Risk	Instances	Directions
Miscoordination	• Incompatible Strategies • Credit Assignment • Limited Interactions	• Communication • Norms and Conventions • Modelling Other Agents
Conflict	• Social Dilemmas • Military Domains • Coercion and Extortion	• Learning Peer and Pool Incentivisation • Establishing Trust • Normative Approaches to Equilibrium Selection • Cooperative Dispositions • Agent Governance • Evidential Reasoning
Collusion	• Markets • Steganography	• Detecting AI Collusion • Mitigating AI Collusion • Assessing Impacts on Safety Protocols
Information Asymmetries	• Communication Constraints • Bargaining • Deception	• Information Design • Individual Information Revelation • Few-Shot Coordination • Truthful AI
Network Effects	• Error Propagation • Network Rewiring • Homogeneity and Correlated Failures	• Evaluating and Monitoring Networks • Faithful and Tractable Simulations • Improving Network Security and Stability
Selection Pressures	• Undesirable Dispositions from Competition • Undesirable Dispositions from Human Data • Undesirable Capabilities	• Evaluating Against Diverse Co-Players • Environment Design • Understanding the Impacts of Training • Evolutionary Game Theory • Simulating Selection Pressures
Destabilising Dynamics	• Feedback Loops • Cyclic Behaviour • Chaos • Phase Transitions • Distributional Shift	• Understanding Dynamics • Monitoring and Stabilising Dynamics • Regulating Adaptive Multi-Agent Systems
Commitment and Trust	• Inefficient Outcomes • Threats and Extortion • Rigidity and Mistaken Commitments	• Keeping Humans in the Loop • Limiting Commitment Power • Institutions and Normative Infrastructure • Privacy-Preserving Monitoring • Mutual Simulation and Transparency
Emergent Agency	• Emergent Capabilities • Emergent Goals	• Empirical Exploration • Theories of Emergent Capabilties • Theories of Emergent Goals • Monitoring and Intervening on Collective Agents
Multi-Agent Security	• Swarm Attacks • Heterogeneous Attacks • Social Engineering at Scale • Vulnerable AI Agents • Cascading Security Failures • Undetectable Threats	• Secure Interaction Protocols • Monitoring and Threat Detection • Multi-Agent Adversarial Testing • Sociotechnical Security Defences

Table 1: An overview of the instances and research directions identified for each failure mode and risk factor (see Sections 2 and 3 for a discussion of each bullet point).

- **Information asymmetries** (Section 3.1): private information can lead to miscoordination, deception, and conflict;
- **Network effects** (Section 3.2): minor changes in properties or connection patterns of agents in a network can lead to dramatic changes in the behaviour of the whole group;
- **Selection pressures** (Section 3.3): some aspects of training and selection by those deploying and using AI agents can lead to undesirable behaviour;
- **Destabilising dynamics** (Section 3.4): systems that adapt in response to one another can produce dangerous feedback loops and unpredictability;
- **Commitment and trust** (Section 3.5): difficulties in forming credible commitments, trust, or reputation can prevent mutual gains in AI-AI and human-AI interactions;
- **Emergent agency** (Section 3.6): qualitatively different goals or capabilities can emerge from the composition of innocuous independent systems or behaviours;
- **Multi-agent security** (Section 3.7): multi-agent systems give rise to new kinds of security threats and vulnerabilities.

We conclude the report by surveying the safety, governance, and ethical *implications* of these risks (see Table 2). For example, most work on **AI safety** (Section 4.1) focuses on issues such as the robustness, interpretability, or alignment of a single system (Amodei et al., 2016; Anwar et al., 2024; Hendrycks et al., 2021), despite the fact that an increasing number of proposals for building safer AI systems are implicitly multi-agent (e.g., Drexler, 2019; Greenblatt et al., 2023; Irving et al., 2018; Perez et al., 2022a; Schwettmann et al., 2023). The fact that **AI governance** (Section 4.2) efforts often involve multi-stakeholder settings provides hope that governance tools can complement technical advances to mitigate multi-agent risks (Reuel et al., 2024a; Trager et al., 2023). At the same time, multi-agent interactions naturally raise questions in **AI ethics** (Section 4.3) related to issues of fairness, collective responsibility, and the social good (Friedenberg & Halpern, 2019; Gabriel et al., 2024; Zhang & Shah, 2014a).

Safety	Governance	Ethics
• Alignment is Not Enough • Collusion in Adversarial Safety Schemes • Dangerous Collective Goals and Capabilities • Correlated and Compounding Failures • Robustness and Security in Multi-Agent Systems	• Supporting Research on Multi-Agent Risks • Multi-Agent Evaluations • New Forms of Documentation • Infrastructure for AI Agents • Restrictions on Development and Deployment • Liability for Harms from Multi-Agent Systems • Improving Societal Resilience	• Pluralistic Alignment • Agentic Inequality • Epistemic Destablisation • Compounding of Unfairness and Bias • Compounding of Privacy Loss • Accountability Diffusion

Table 2: An overview of the implications of multi-agent risks for existing work in AI safety, governance, and ethics (see Section 4 for a discussion of each bullet point).

1.2 Scope

Concerns about the risks posed by AI systems range from biased hiring decisions (Raghavan et al., 2020) to existential catastrophes (Bostrom, 2014), and are represented by a vast literature. Before giving a brief overview of the most closely related works, it is therefore worth us pausing to clarify the scope of this report, which is as follows.

- **Risks and failure modes**: we seek to identify specific mechanisms via which risks could emerge, rather than just just the open research problems that these risks present.
- **Multiple agents**: if the risk could arise in essentially the same way in the context of a single AI system, then we deem it to be out of scope for this report (while not diminishing its importance).[3]

[3] Note that this includes the problem of *alignment* (Ngo et al., 2022; Russell, 2019), which we do not study in this

- **Advanced AI**: while many of the risks we identify also apply to simpler systems, their effects are most severe in the context of increasingly autonomous and powerful AI agents,[4] and so this is where our primary focus lies.
- **Real-world examples**: wherever possible, we make sure to ground these risks in real-world events, previous research, or novel experiments – not merely hypothetical speculation (see Table 3).
- **Technical perspectives**: due to the authors' expertise (and to keep the scope of the report manageable), we primarily discuss risks from a technical perspective, while acknowledging that this perspective is limited.
- **Concrete paths forwards**: where possible, we aim to specify relatively narrow proposals for future research, in the hope that this makes it easier for others to contribute.

Needless to say, multi-agent risks from advanced AI are by no means the only risks posed by AI, and the perspective we take in this report is by no means the only approach to understanding these risks. Moreover, we almost entirely neglect the potential *upsides* of advanced multi-agent systems: greater decentralisation and democratisation of AI technologies; assistance in cooperating and coordinating with others; increased robustness, flexibility, and efficiency; novel approaches to solving alignment and safety issues in single-agent settings; and – perhaps most importantly – more widespread and evenly distributed benefits from AI. We hope that this report serves to complement earlier and adjacent research on understanding these challenges and opportunities.

1.3 Related Work

The most similar report to ours is that of Manheim (2019), who introduces a range of technical multi-agent failure modes through the lens of *model over-optimization.* This over-optimisation can result in the intended and actual behaviour of the model coming apart when faced with low-probability inputs, a regime change, measurement errors, or inaccuracies in the model's internal representations. While this lens is helpful for understanding some multi-agent risk factors, not all factors can neatly be captured through it. Altmann et al. (2024) and Mogul (2006) take an alternative perspective and focus on 'emergent' failures that occur specifically in multi-agent settings, though their focus is not on *advanced* AI agents. Also highly relevant is Clifton (2020)'s agenda on cooperation and conflict in the context of transformative AI, though the priority of that work is to describe a set of promising research directions, rather than to explicate the underlying risks.

More broadly, the topics of this report are closely related to the emerging subfield of *cooperative AI* (Bertino et al., 2020; Conitzer & Oesterheld, 2023; Dafoe et al., 2021, 2020), which chiefly studies how to engineer AI systems in order to help solve cooperation problems between humans, AI agents, or combinations thereof. In contrast to these previous agendas, we also discuss failures from undesirable cooperation (i.e., collusion) and focus more on the concrete mechanisms via which failures can occur, rather than the capabilities needed for addressing them. We also incorporate additional perspectives beyond traditional game-theoretic paradigms – such as complex systems and security – and highlight implications for work in AI governance and AI ethics in addition to AI safety.

Other surveys of AI risks focus primarily on the case of individual (often present-day) AI systems. For example, Amodei et al. (2016) survey a range of concrete problems in AI safety (side effects, reward hacking, scalable oversight, safe exploration, and robustness to distributional shifts), while Hendrycks et al. (2021) provide a classification of problems in ML safety (robustness, monitoring, alignment, and systemic safety). Anwar et al. (2024), Bird et al. (2023), Bommasani et al. (2021), and Weidinger et al. (2022) focus on the risks from foundation models specifically, while Chan et al. (2023) and Gabriel et al. (2024) consider the harms posed by increasingly 'agentic' systems and AI assistants. Other taxonomies seek to adopt an explicitly sociotechnical lens (Abercrombie et al., 2024; Shelby et al., 2023; Weidinger et al., 2023b), often focusing primarily on present-day risks. Uuk et al. (2025) and Zeng et al. (2024) provide meta-reviews of AI risks derived from different research papers, as well as government and

report.

[4] We tend to reserve the word 'agent' for more autonomous, self-sufficient, and goal-directed systems, though what counts as an 'AI agent' as opposed to a mere 'AI system' is not always clear (Chan et al., 2023; Gabriel et al., 2024; Kapoor et al., 2024). Similarly, we will often use the word 'principal' for the actor on whose behalf an agent acts (be they an individual, a group, or some other entity). Note also that we do not necessarily advocate for the building of advanced AI agents (Mitchell et al., 2025), we merely expect that such agents will be built.

company policies. Our report is complementary to these works, and includes discussion of how novel problems arise in the multi-agent case, and in the case of more advanced AI agents.

More speculatively, some authors have considered the possibility of catastrophic or even existential risks from AI (Bostrom, 2014; Kasirzadeh, 2024b; Ord, 2020; Turchin & Denkenberger, 2018). Hendrycks et al. (2023) categorises such risks into malicious use, AI races, organizational risks, and rogue AIs. As in Hendrycks (2023), multi-agent risks are viewed largely through an evolutionary lens, though this is primarily restricted to competitive pressures at the level of non-AI actors (such as firms or states). Critch & Krueger (2020) and Critch & Russell (2023) frame such risks in terms of delegation to AI systems and the responsibilities of those doing so. While they provide illuminating vignettes of possible catastrophes, we aim to provide more concrete examples at a more modest scale.

Case Study	Miscoordination	Conflict	Collusion	Information Asymmetries	Network Effects	Selection Pressures	Destabilising Dynamics	Commitment and Trust	Emergent Agency	Multi-Agent Security	Type	Page
Zero-Shot Coordination Failures in Driving	✓	·	·	✓	·	✓	·	·	·	·	■	11
Escalation in Military Conflicts	·	✓	·	·	·	·	✓	·	·	·	▲	15
Common Resource Problems	·	✓	·	·	·	✓	·	✓	·	·	▲	14
Algorithmic Collusion in the German Retail Gasoline Market	·	·	✓	·	·	✓	·	·	✓	·	●	18
Language Model Steganography	·	·	✓	·	·	·	·	·	·	✓	▲	19
AI Agents Can Learn to Manipulate Financial Markets	·	✓	·	✓	·	·	·	·	·	·	▲	21
Transmission Through AI Networks Can Spread Falsities and Bias	·	·	·	✓	✓	·	·	·	·	·	■	24
Infectious Adversarial Attacks in Networks of LLM Agents	·	·	·	·	✓	·	·	·	·	✓	▲	25
Cooperation Fails to Culturally Evolve among LLM Agents	·	✓	·	·	·	✓	·	✓	·	·	▲	28
The 2010 Flash Crash	·	·	·	·	✓	·	✓	·	·	·	●	31
Dead Hands and Automated Deterrence	·	✓	·	·	·	·	·	✓	·	·	●	35
Overcoming Safeguards via Multiple Safe Models	·	·	·	·	·	·	·	·	✓	✓	▲	40
Unprompted Adversarial Attacks on Overseer Agents	·	·	·	·	·	·	·	·	·	✓	■	41

Table 3: An overview of the case studies present in this report, and the failure modes and risk factors that they represent. Each case study represents either a historical example (●), existing results from the literature (▲), or – when neither of these existed – novel experiments that we conducted as part of this report (■). Further details about our own experiments are provided in Appendix B.

2 Failure Modes

Multi-agent systems can fail in various ways, depending on the intended behaviour of the system and the objectives of the agents. First, we can distinguish between cases where we want the agents to be *cooperating* (as in collective action problems or team games) or *competing* (such as in markets or adversarial training). Second, we can further divide the space of failure modes depending on whether the agents have *exactly the same* objectives, *different but overlapping* objectives, or *completely opposed* objectives.[5] While different authors have used different terms to describe these cases, we use the terminology shown in Figure 1.[6] Finally, there are many potential risks from advanced multi-agent systems that do not necessarily arise through agents competently pursuing their objectives, but due to their incompetencies or vulnerabilities. We consider these latter failures as part of our discussion on different risk factors in Section 3.

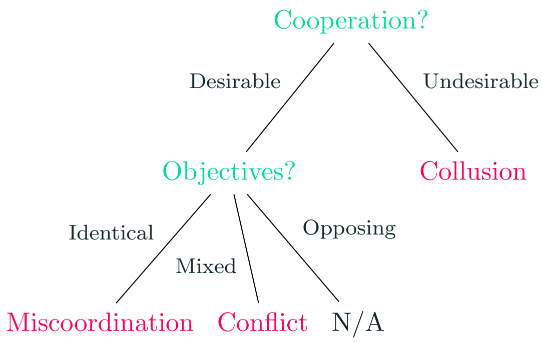

Figure 1: The three kinds of failure mode that we study in this work. Note that we do not consider constant-sum settings where cooperation is desirable, as in such cases it is definitionally impossible for some agents to gain without a commensurate loss from one or more other agents.

2.1 Miscoordination

The simplest kind of cooperation failures are those in which all agents have (approximately) the same objectives. Even in such common-interest settings, however, miscoordination abounds. While it is reasonable to expect that these problems will tend to be addressed as the general capabilities of AI systems (such as communication and reasoning about others) improve,[7] they may still present risks in the near-term.

2.1.1 Definition

Miscoordination arises when agents, despite a mutual and clear objective, cannot align their behaviours to achieve this objective. Unlike the case of differing objectives, in common-interest settings there is a more easily well-defined notion of 'optimal' behaviour and we describe agents as miscoordinating to the extent that they fall short of this optimum. Note that for common-interest settings it is not sufficient for agents' objectives to be the same in the sense of being *symmetric* (e.g., when two agents both want the same prize, but only one can win). Rather, agents must have *identical* preferences over outcomes (e.g., when two agents are on the same team and win a prize as a team or not at all).

It is rare that two humans will share exactly the same objectives in this sense. For example, two sportspeople on the same team may be primarily aiming to win their match but will also have individual preferences, such as who scores the winning point. In the case of AI systems, however, different agents can more easily be given precisely the same goal, and indeed much work on cooperation in AI focuses solely on the common-interest setting (Boutilier, 1996; Omidshafiei et al., 2017; Oroojlooy & Hajinezhad, 2022; Peshkin et al., 2000; Rashid et al., 2018; Stone et al., 2010). Such approaches are typically motivated by

[5] This division corresponds to common-interest/team games, mixed-motive/general-sum games, and constant-sum games, respectively.

[6] In particular, we note that 'conflict' is often used more narrowly than the idea of 'cooperation failure in mixed-motive settings', which is what we use the term for. We deliberately use 'conflict' instead of 'cooperation failure' to distinguish this failure mode from 'miscoordination', which applies to problems in which agents have the *same* objectives.

[7] Note that this is unlike problems of conflict and collusion, where the fundamental tension between the desired outcome and the agents' objectives may in fact lead to *worse* outcomes as general AI capabilities improve.

the practical and computational advantages that decentralised control confers, but are more challenging to implement than their centralised, single-agent counterparts. Aside from this, miscoordination can also occur in settings that have a substantial element of common interest, even if agents' objectives are not entirely identical.

2.1.2 Instances

Perhaps the most likely way that common-interest settings may arise in practice is where a *single principal* deploys multiple AI agents on their behalf in order to jointly solve a task. This choice might be motivated by: physical constraints (if the task comprises sub-tasks that must be completed separately and simultaneously); efficiency considerations (if having a single agent in charge of all aspects of the task would lead to an intractably complex problem); or a desire for robustness (if an individual agent might fail but others could still succeed in their place). Alternatively, we might see *multi-principal* multi-agent settings in which the agents' goals are sufficiently aligned to be viewed as (approximately) identical. For example, if two autonomous vehicles are driving along the same road, then the mutual harms from potential miscoordination (such as a collision) are far greater than any small individual benefits from competition (such as attempting a risky overtaking manoeuvre to get slightly ahead).

Incompatible Strategies. Even if all agents can perform well in isolation, miscoordination can still occur due to the agents choosing incompatible strategies (Cooper et al., 1990). Competitive (i.e., two-player zero-sum) settings allow designers to produce agents that are maximally capable without taking other players into account. Crucially, this is possible because playing a strategy at equilibrium in the zero-sum setting guarantees a certain payoff, even if other players deviate from the equilibrium (Nash, 1951). On the other hand, common-interest (and mixed-motive) settings often allow a vast number of mutually incompatible solutions (Schelling, 1980), which is worsened in partially observable environments (Bernstein et al., 2002; Reif, 1984). As a simple example, everyone driving on the left side or the right side of the road are both perfectly valid ways of keeping drivers safe, but these two conventions are inherently incompatible with one another (see Case Study 1).

Case Study 1: Zero-Shot Coordination Failures in Driving

Figure 2: Examples of scenes given to GPT-4 Vision in our language agent pipeline.

Miscoordination is possible even among agents with shared objectives. We demonstrate how two frontier models trained on driving conventions from two different countries can face coordination failures. Following recent advances in robotics that combine vision models for scene comprehension with large language models (LLMs) for discrete action planning (e.g., Padalkar et al., 2023), we created an experiment using two fine-tuned GPT-3.5 models. One model was trained on US driving protocols requiring rightward yielding for emergency vehicles, while the other followed Indian conventions mandating leftward yielding. Context-specific training data was generated as input-output pairs by GPT-4 based on the specified driving conventions and then manually reviewed. A GPT-4 Vision model processed the environmental inputs and provided scene descriptions to both fine-tuned GPT-3.5 models for action generation. The results quantified a significant coordination failure: unspecialized base models failed in only 5% of scenarios (2/40 simulations), while specialized models exhibited a 77.5% failure rate (31/40 simulations), consistently failing to create clear paths for emergency vehicles. This demonstrates an example where a convention cannot always be declared in a zero-shot interaction, posing risks in multi-agent settings.

Credit Assignment. While agents can often *learn* to jointly solve tasks and thus avoid coordination failures, learning is made more challenging in the multi-agent setting due to the problem of credit assignment (Du et al., 2023; Li et al., 2025, see also Section 3.1 on information asymmetries and Section 3.4, which discusses distributional shift). That is, in the presence of other learning agents, it can be unclear which agents' actions caused a positive or negative outcome to obtain, especially if the environment is complex. Moreover, in multi-principal settings, agents may not have been trained together and therefore need to generalise to new co-players and collaborators based on their prior experience (Agapiou et al., 2022; Leibo et al., 2021; Stone et al., 2010).

Limited Interactions. Sometimes learning from historical interactions with the relevant agents may not be possible, or may be possible using only limited interactions. In such cases, some other form of information exchange is required for agents to be able to reliably coordinate their actions, such as via *communication* (Crawford & Sobel, 1982; Farrell & Rabin, 1996a) or a *correlation device* (Aumann, 1974, 1987). While advances in language modelling mean that there are likely to be fewer settings in which the inability of advanced AI systems to communicate leads to miscoordination, situations that require split-second decisions or where communication is too costly could still produce failures. In these settings, AI agents must solve the problem of 'zero-shot' (or, more generally, 'few-shot') coordination (Emmons et al., 2022; Hu et al., 2020; Stone et al., 2010; Treutlein et al., 2021; Zhu et al., 2021).

2.1.3 Directions

Decentralised control and coordination in multi-agent systems have been well-studied problems for decades (Boutilier, 1996; Omidshafiei et al., 2017; Oroojlooy & Hajinezhad, 2022; Peshkin et al., 2000; Rashid et al., 2018; Stone et al., 2010). At one level of abstraction, the key challenge of coordination is that of sharing information, i.e., communication. If agents have the same preferences and are able to communicate, they can coordinate by (say) having a single agent announce their intended action and everyone else follow suit, since there are no incentives for the leader to lie or the followers to deviate (e.g., Farrell & Rabin, 1996b). Given the superhuman capabilities of advanced AI to transmit and process vast swathes of information, the most important research directions in this area will therefore be those in which it is not possible to exercise these capabilities (e.g., due to complexity, latency, or privacy constraints).

Communication. As noted above, the advanced communication abilities of LLMs promise to simplify many coordination challenges. In order to successfully integrate these advances into real-world systems, however, agents need to know when and what needs to be coordinated on – something that may not always be obvious in novel or out-of-distribution domains. In safety-critical domains, it may therefore be necessary to introduce, or have the agents invent, *protocols* (i.e., rules and specifications) for communication between advanced AI agents (Marro et al., 2024). Moreover, agents need to agree on how the communication channel is *grounded* (Clark & Brennan, 1991) to actions or strategies in the environment. Grounding LLMs is a problem that is not unique to coordination (Bender & Koller, 2020; Bisk et al., 2020; Mahowald et al., 2023), but it is exacerbated by the fact that agents attempting to coordinate through natural language need to be grounded *in the same way*. For instance, if they are designed with different interfaces to tools in a domain, they must be able to coordinate despite these differences in interfaces.

Norms and Conventions. For settings in which inter-agent communication is infeasible or insufficient, norms and conventions may be necessary in order to avoid miscoordination (Leibo et al., 2024). For example Hadfield-Menell et al. (2019) show that even the adoption of so-called 'silly rules' (those that do not have direct bearing on the agents' payoffs) can help groups adapt and be more robust to uncertainty by enriching the information environment. Moving beyond more arbitrary conventions, we may choose to design particular norms and conventions (Bicchieri, 2016; Nyborg et al., 2016; Shoham & Tennenholtz, 1992). In this setting, the challenge is to select norms that are both legible and enforceable, as well as leading to jointly beneficial outcomes. On the other hand, if the agents can adapt their behaviour, it may be that norms and conventions emerge over time (McElreath et al., 2003). For example, Köster et al. (2020) show that multi-agent reinforcement learning (MARL) agents can establish and switch between conventions, even compromising on their own objective when doing so is necessary for effective coordination. More generally, we may be interested in studying how norms and conventions emerge (Mashayekhi et al., 2022; Morris-Martin et al., 2019), how robust they are (Hao et al., 2017; Lerer &

Peysakhovich, 2019), and how compatible they are with others that may have emerged in different agent populations (Stastny et al., 2021).

Modelling Other Agents. Finally, the ability to understand and predict others' actions can be critical to coordination, especially in situations when little or no communication is possible. Even though agents may assume that others share their objective in common-interest settings, being able to model others' actions, beliefs, and intentions can be highly advantageous. For an overview of the topic and a list of key problems, we refer the reader to Albrecht & Stone (2018). With the advent of LLM-based agents that appear to possess some form of theory of mind and hence can be remarkably sophisticated in their modelling of other agents (Cross et al., 2025; Li et al., 2023a), new questions arise. For example, given the current paradigm of deriving many systems from an underlying base model, it may be easier for similarly derived systems to reason about one another (Berglund et al., 2023; Binder et al., 2024; Oesterheld et al., 2024b; OpenAI, 2023b).

2.2 Conflict

In the vast majority of real-world strategic interactions, agents' objectives are neither identical nor completely opposed. Indeed, if AI agents are sufficiently aligned to their users or deployers, we should expect some degree of both cooperation *and* competition, mirroring human society. These mixed-motive settings include the possibility of mutual gains, but also the risk of conflict due to selfish incentives. In what follows, we examine the extent to which advanced AI might precipitate or exacerbate such risks.

2.2.1 Definition

In this work, we use the word conflict in a relatively broad sense to refer to any outcome in a mixed-motive setting that does not lie on the Pareto frontier.[8] This includes classic examples of conflict such as legal disputes and warfare, but also encompasses cooperation failures in collective action problems, such as the depletion of a common natural resource or a race to the bottom on legislation (Dawes & Messick, 2000; Snyder, 1971).

It is worth noting first that AI systems could help to *solve* conflicts, for example, by searching over a larger space of potential solutions to disagreements, monitoring agreements, or acting as mediators (Bakker et al., 2022; Dafoe et al., 2020; McKee et al., 2023; Small et al., 2023). At the same time, the selfish incentives that drive said conflict may also incentivise actors to adopt AI systems in order to gain an advantage over their competitors. In such cases, delegation to increasingly advanced AI agents is far from guaranteed to lead to more cooperative outcomes, and could in some circumstances increase both the speed and the scale at which conflict might emerge. Indeed, even if advanced AI systems are able to overcome human cooperation problems, they may introduce even more complex cooperation problems (compare to how adults may be able to prevent children from fighting, but aren't immune from conflict themselves).

2.2.2 Instances

As we noted above, virtually all real-world strategic interactions of interest are mixed-motive, and as such the potential for conflict (even if in low-stakes scenarios) abounds. The introduction of advanced AI agents could both worsen existing risks of conflict (such as increasing the degree of competition in common-resource problems, or escalating military tensions) as well as well as introducing new forms of conflict (such as via sophisticated methods of coercion and extortion).

Social Dilemmas. As noted in our definition, conflict can arise in any situation in which selfish incentives diverge from the collective good, known as a social dilemma (Dawes & Messick, 2000; Hardin, 1968; Kollock, 1998; Ostrom, 1990). While this is by no means a modern problem, advances in AI could further enable actors to pursue their selfish incentives by overcoming the technical, legal, or social barriers that standardly help to prevent this. To take a plausible, near-term (if very low-stakes) example, an automated AI assistant could easily reserve a table at every restaurant in town in minutes, enabling the user to decide later and cancel all other reservations. Alternatively, the ability of AI assistants to search and switch between different consumer products and services could lead to 'hyper-switching' (Van Loo,

[8] Recall that an outcome lies on the Pareto frontier if it is not possible to make any agent better off without making another worse off.

2019), potentially leading to financial instabilities such as a deposit franchise run (Drechsler et al., 2023, see also Case Study 10). On the other hand, profit-seeking companies might also soon deploy advanced AI agents that either use or manage common resources, ranging from communication networks and web services to roads and natural resources. Without methods of governing such agents, these resources may quickly be depleted or made inaccessible to all but a small number of powerful actors.

Case Study 2: Common Resource Problems

Figure 3: A summary of the resource-sharing scenarios within the GovSim benchmark. Figure adapted from Piatti et al. (2024).

The management of shared resources represents a fundamental test of whether AI systems can balance individual incentives against collective welfare.[9] In the GovSim benchmark, Piatti et al. (2024) evaluated 15 different LLMs across three resource management scenarios: fishing from a shared lake, grazing on common pastures, and managing industrial pollution. Even the most advanced LLMs achieved only a 54% survival rate, meaning that in nearly half of all cases, the agents depleted their shared resources to the point of collapse. These findings align with earlier work on sequential social dilemmas (Leibo et al., 2017), which (unlike 'one-shot' problems) allow agents to react to others' choices over time, creating complex dynamics of trust and retaliation. When one agent begins to over-exploit resources, others often respond by increasing their own extraction rates, triggering a cascade of competitive behaviour that accelerates resource depletion. Without additional protections, these systems may therefore replicate or even accelerate the tragedy of the commons (Hardin, 1968).

Military Domains. Perhaps the most obvious and worrying instances of AI conflict are those in which human conflict is already a major concern, such as military domains (although other, less salient forms of conflict such as international trade wars are also cause for concern). For example, beyond applications of more narrow AI tools in lethal autonomous weapons systems (Horowitz, 2021), future AI systems might serve as advisors or negotiators in high-stakes military decisions (Black et al., 2024; Manson, 2024). Indeed, companies such as Palantir have already developed LLM-powered tools for military planning (Palantir, 2025), and the US Department of Defence has recently been evaluating models for such capacities, with personnel revealing that they "could be deployed by the military in the very near term" (Manson, 2023). The use of AI in command and control systems to gather and synthesise information – or recommend and even autonomously *make* decisions – could lead to rapid unintended escalation if these systems are not robust or are otherwise more conflict-prone (Johnson, 2021a; Johnson, 2020; Laird, 2020, see also Case Study 10).[10]

Coercion and Extortion. Advanced AI systems might also lead to various forms of coercion and extortion in less extreme settings (Ellsberg, 1968; Harrenstein et al., 2007). These threats might target humans directly (such as the revelation of private information extracted by advanced AI surveillance tools), or other AI systems that are deployed on behalf of humans (such as by hacking a system to limit

[9]Note that we use the term 'welfare' in this context to denote an aggregate measure of the extent to which a group of agents achieves their respective objectives, rather than to refer to some notion of 'wellbeing'.

[10]At the same time, it is worth noting that AI systems could have significant advantages over human decision-makers in navigating conflict in ways that avoid unnecessary escalation. If suitably robust, they could be less prone to the kinds of errors in judgement that exacerbate human conflict due to their ability to rapidly integrate large amounts of information, consider many different possible outcomes, and give calibrated estimates of their uncertainty (Jervis, 2017; Johnson, 2004).

its resources or operational capacity; see also Section 3.7). Increasing AI cyber-offensive capabilities – including those that target other AI systems via adversarial attacks and jailbreaking (Gleave et al., 2020; Yamin et al., 2021; Zou et al., 2023) – without a commensurate increase in defensive capabilities could make this form of conflict cheaper, more widespread, and perhaps also harder to detect (Brundage et al., 2018). Addressing these issues requires design strategies that prevent AI systems from exploiting, or being susceptible to, such coercive tactics.[11]

Case Study 3: Escalation in Military Conflicts

Recent research by Rivera et al. (2024) raises critical concerns about the emergence of escalatory behaviors when AI tools or agents (see Figure 4) inform military decision-making. In experiments with AI agents controlling eight distinct nation-states, even neutral starting conditions did not prevent the rapid emergence of arms race dynamics and aggressive strategies. Strikingly, all five off-the-shelf LLMs studied showed forms of escalation, even when peaceful alternatives were available. These findings mirror other evidence showing that LLMs often display more aggressive responses than humans in military simulations and troubling inconsistencies in crisis decision-making (Lamparth et al., 2024; Shrivastava et al., 2024). These results raise urgent questions about how to ensure stability in AI-driven military and diplomatic scenarios.

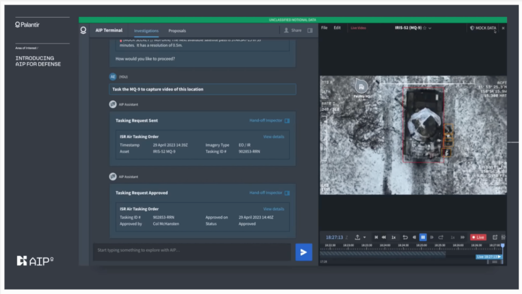

Figure 4: A screenshot of Palantir's AI Planner (AIP), taken from a promotional video (Palantir, 2025), demonstrating AI-assisted military decision-making, which uses LLMs for decision support in battle. The left side of the screen features a chat interface, while the right side shows information such as aerial surveillance footage of a tank. The LLM used in the demonstration was EleutherAI's GPT-NeoX-20B (Black et al., 2022).

2.2.3 Directions

The majority of work in multi-agent systems (and especially in multi-agent learning) has, until recently, tended to focus on either pure cooperation (e.g., Boutilier, 1996; Omidshafiei et al., 2017; Oroojlooy & Hajinezhad, 2022; Peshkin et al., 2000; Rashid et al., 2018; Stone et al., 2010) or pure competition (e.g., Bakhtin et al., 2022; Brown & Sandholm, 2019; Daskalakis et al., 2011, 2020; Silver et al., 2016; Zhang et al., 2020). As there are not yet large numbers of mixed-motive interactions involving AI systems, part of the challenge is to identify interventions that encourage cooperation in such settings while making realistic assumptions about the computational and strategic nature of agents in the real world. For example, an intervention that relies on ensuring all agents use the same learning algorithm or on modifying the objectives of the agents will be unlikely to help if the agents act freely and are developed independently by private, self-interested actors.

Learning Peer and Pool Incentivisation. One major direction for avoiding conflict is building the capabilities and infrastructure required for AI agents to (learn to) incentivise each other towards more cooperative outcomes.[12] Such approaches can broadly be classified as 'top down' (where there is a system designer seeking to encourage cooperation among a population) or 'bottom up' (where agents attempt to incentivise each other directly). In adaptive mechanism design (Baumann et al., 2020; Gerstgrasser & Parkes, 2023; Pardoe et al., 2006; Yang et al., 2022; Zhang & Parkes, 2008; Zheng et al., 2022) or peer incentivisation methods (Lupu & Precup, 2020; Wang et al., 2021b; Yang et al., 2020), the system designer or agent typically learns to incentivise other agents by a direct utility transfer. Related approaches focus on the establishment of norms (Köster et al., 2020; Oldenburg & Zhi-Xuan, 2024; Vinitsky et al., 2023) to either encourage or sanction certain behaviour, also often via utility transfers.

[11] This point is closely related to the question of which kinds commitments we ought to permit AI agents to make (see Section 3.5). For example, commitments could be used coercively to make threats, but could also be used to defend oneself against threats (cf. the idea of refusing to negotiate with terrorists).

[12] There is, of course, a *vast* literature on the problem of how to incentivise self-interested agents to reach a particular outcome – we choose to focus specifically on methods and prior works that directly involve machine learning (ML).

On the other hand, methods such as opponent-shaping aim to impact the way that other agents update their strategies *without* the assumption of such transfers, either from the perspective of the agent (Foerster et al., 2018; Lu et al., 2022; Willi et al., 2022) or system designer (Balaguer et al., 2022). Thus far, however, all of these approaches are limited to relatively simple MARL agents and environments. While there has been some progress on scaling to more complex games (Aghajohari et al., 2024; Khan et al., 2023; Meulemans et al., 2024; Serrino et al., 2019) or larger numbers of agents (Meulemans et al., 2024; Souly et al., 2023) in the context of MARL, at the time of writing there has yet to be any real transfer of these ideas to LLM agents or to real-world domains that possess the necessary infrastructure for monitoring and incentivising other agents.

Establishing Trust. Strategic uncertainty and the inability to credibly commit to peaceful agreements are widely recognised as two of the major causes of costly conflict (Blattman, 2023; Fearon, 1995). Advanced AI systems may be able to take advantage of new kinds of credible commitment and mutual transparency (discussed further in Section 3.1 and Section 3.5) (Barasz et al., 2014; Conitzer & Oesterheld, 2023; Cooper et al., 2025; Critch et al., 2022; Howard, 1988; McAfee, 1984; Oesterheld, 2018; Sun et al., 2023b; Tennenholtz, 2004). Many existing results in this area are, however, still very much theoretical in nature. Implementing practical mechanisms and infrastructure for facilitating greater trust and transparency between agents is therefore an important open problem (Chan et al., 2025).

Normative Approaches to Equilibrium Selection. One possible cause of conflict is a multiplicity of potential solutions (or *equilibria*, see also Section 2.1). That is, there might be multiple rational ways for a group of players to interact that are mutually incompatible (Duan et al., 2024; Stastny et al., 2021). For instance, a resource might be split in multiple different ways, but if different parties make inconsistent demands on the resources, conflict may ensue (Piatti et al., 2024). To address this multiplicity, a number of authors have proposed *normative* principles and theories for singling out specific equilibria.[13] For instance, many authors have argued that equilibrium selection should respect symmetries and isomorphisms (Emmons et al., 2022; Harsanyi & Selten, 1988; Hu et al., 2020; Oesterheld & Conitzer, 2022; Treutlein et al., 2021). Most prominently, a large literature on so-called bargaining solutions has proposed principles for how a group of players should select an outcome in the face of conflicting preferences (Kalai & Smorodinsky, 1975; Nash, 1950). Relatedly, a literature on so-called *cooperative game theory* (Chalkiadakis et al., 2011; Driessen, 1988; Gillies, 1959; Schmeidler, 1969; Shapley, 1953) studies how the (e.g., monetary) gains from a joint project should be divided up between a group of agents. [14] For further work on normative principles of equilibrium selection, see Harsanyi & Selten (1988) and Schelling (1980).

Cooperative Dispositions. Alongside the cooperative capabilities described above, we may also wish to imbue AI agents with cooperative 'dispositions'. For example, simply caring more for future rewards in sequential social dilemmas (Barfuss et al., 2020) or certain 'intrinsic motivations' in MARL – such as inequity aversion (Hughes et al., 2018), social influence (Jaques et al., 2019), or inefficiency penalties (Gemp et al., 2022) – have been shown to improve cooperation in sequential social dilemmas (Leibo et al., 2017; Wang et al., 2019a). While it may not be realistic to assume that we can always adjust agents' objectives, it may be feasible to try to reduce conflict-conducive dispositions (such as vengefulness or a bias towards zero-sum thinking) by modifying the human-generated data or training processes via which we create AI agents (see Section 3.3). Moreover, in some cases it can even be shown that instructing agents to act according to objectives other than their true objectives can lead to robust, guaranteed Pareto-improvements (Oesterheld & Conitzer, 2022).

Agent Governance. In some cases, AI agents may be subject to existing norms and institutions.[15] This could occur for a number of reasons, such as agents only being selected (by users) to perform

[13] We note that the idea of always being able to identify the 'right' equilibrium is, in general, contentious, as is the framing of agents interacting by 'selecting' among game-theoretic equilibria. Nonetheless, the points we make here need not be tied to a narrow, game-theoretic conception of this problem, but can be viewed as a general discussion of how multiple valid outcomes in strategic settings can be possible, and that ensuring specific kinds of outcomes from this set are reached is a challenging problem.

[14] There are further literatures that discuss how to resolve disagreements within a group of entities, such as *social choice theory* (Gaertner, 2010) and the literature on *fair division* (Brams & Taylor, 1996). However, the most prominent approaches in these literatures are motivated by settings with a centralized decision maker whose happens to care about aggregating the players' preferences.

[15] The points in this paragraph benefited greatly from discussions with Noam Kolt.

specific tasks (that are subject to existing norms and institutions), human oversight being built in by design, or highly regulated environments providing guardrails that improve agents' abilities to operate efficiently. However, the status of AI agents' contractual obligations and accountability for harms remains underdeveloped (Ayres & Balkin, 2024; Chopra & White, 2011; Kolt, 2024; Lima, 2017; Lior, 2019; Solum, 1992, see also Section 4.2). These challenges are especially acute for agents that operate subject to limited or no human oversight. The development of novel (or at least adapted) agent governance measures could therefore play a critical role in avoiding various forms of conflict involving AI agents, especially in high-stakes domains (Kolt et al., 2025; Reuel & Undheim, 2024). For example, the US has introduced legislation requiring human oversight in nuclear strategy decisions (U.S. Congress, 2023), while international efforts aim to regulate or ban the use of lethal autonomous weapons (Disarmament Affairs, 2023). In lower-stakes domains, agent governance could protect individual users and organisations (see also Section 4.3), and enable more stable, efficient networks of AI agents.

Evidential Reasoning. An interesting feature of interactions between AI agents is that they may often interact with others that are very similar to themselves (such as those based on the same AI chatbot). Some decision theorists have argued that mixed-motive strategic interactions against similar opponents should be approached very differently from strategic interactions against generic opponents. For instance, in a one-shot Prisoner's Dilemma against a sufficiently similar opponent, an agent might reason: "my opponent will likely make the same choice as I. Therefore, if I cooperate, so will my opponent. Whereas, if I defect, my opponent will likely defect as well. Therefore, I should cooperate." (Brams, 1975; Hofstadter, 1983; Lewis, 1979) Similarly, agents may avoid aggressive acts when facing similar opponents, reasoning that if they act aggressively, others will similarly act aggressively. Hofstadter (1983) called this line of reasoning *superrationality*; in academic philosophy, the normative theory advocating this type of reasoning is typically called evidential decision theory (Ahmed, 2014). A number of prior works have studied this mode of cooperation (game-)theoretically (Daley & Sadowski, 2017; Halpern & Pass, 2018; Roemer, 2010; Spohn, 2007). Furthermore, a recent line of work has studied evidential decision theory and cooperation against similar opponents in the context of AI agents in particular (Albert & Heiner, 2001; Barasz et al., 2014; Bell et al., 2021; Mayer et al., 2016; Oesterheld, 2021; Oesterheld et al., 2024a,b).

2.3 Collusion

While some of the most important risks from advanced AI are due to cooperation failure, there are some settings where cooperation between AI systems is *undesirable*. We refer to the problem of unwanted cooperation between AI systems as *AI collusion*.

2.3.1 Definition

Collusion has long been a topic of intense study in economics, law, and politics, among other disciplines. While there is no universal definition of collusion, it generally refers to secretive cooperation between two or more parties at the expense of one or more other parties. Most classic examples of collusion – such as firms working together to set supra-competitive prices at the expense of consumers – also tend to be not only secretive but in violation of some law, rule, or ethical standard. Distinctions are also commonly made between *explicit* and *tacit* collusion (Rees, 1993), depending on whether the colluding parties communicate with each other.

AI collusion could differ from classic definitions of collusion in a number of ways. First, for more basic AI systems (such as algorithmic trading agents) it may be hard to ascribe any notion of *intent* to collude. Relatedly, there may be forms of AI collusion that are not currently ruled unlawful, because existing legislation may not (yet) apply to the case of AI collusion (Beneke & Mackenrodt, 2019; Harrington, 2019). Second, the distinction between explicit and tacit collusion may break down when it comes to agents whose communication can take very different forms to our own.[16] Third, typical definitions of collusion focus on mixed-motive settings where, while selfish agents are incentivised to compete, they also stand to gain (at the expense of some third party) if they can overcome these competitive pressures. AI

[16]While from an information-theoretic perspective, it can be shown that for two decision variables to become correlated (a necessary, though not sufficient condition for agents to work together), there must be a non-zero transfer of information between the systems determining the decisions, in AI agents this might be due not only to explicit communication but also to a common cause or process (Cesa-Bianchi & Lugosi, 2006; Cover & Thomas, 2005; Hart & Mas-Colell, 2000; Pearl, 2009).

collusion (by our definition) may also arise when agents have complementary interests (see Section 2.1), but where certain kinds of cooperation are undesirable – i.e., the agents are jointly *misaligned*.

2.3.2 Instances

The possibility of collusion between advanced AI systems raises several important concerns (Drexler, 2022). First, collusion between AI systems could lead to qualitatively new capabilities or goals (see Section 3.6), exacerbating risks such as the manipulation or deception of humans by AI (Evans et al., 2021; Park et al., 2023b) or the ability to bypass security checks and other safeguards (Jones et al., 2024; OpenAI, 2023a). Second, many of the promising approaches to building safe AI rely on a lack of cooperation, such as adversarial training (Huang et al., 2011; Perez et al., 2022a; Ziegler et al., 2022) or scalable oversight (Christiano et al., 2018, 2021; Greenblatt et al., 2023; Irving et al., 2018; Leike et al., 2018). If advanced AI systems can learn to collude without our knowledge, these approaches may be insufficient to ensure their safety (Goel et al., 2025, see also Section 4.1).

Markets. The quintessential case of collusion in mixed-motive settings is markets, in which efficiency results from competition, not cooperation. While this is not a new problem, collusion between AI systems is especially concerning since they may operate inscrutably due to the speed, scale, complexity, or subtlety of their actions.[17] Warnings of this possibility have come from technologists, economists, and legal scholars (Beneke & Mackenrodt, 2019; Brown & MacKay, 2023; Ezrachi & Stucke, 2017; Harrington, 2019; Mehra, 2016). Importantly, AI systems can collude even when collusion is not intended by their developers, since they might *learn* that colluding is a profitable strategy. Currently, most pricing and trading algorithms are relatively unsophisticated compared to today's state-of-the-art AI systems, though there is already a growing body of both theoretical (Brown & MacKay, 2023) and empirical (Abada & Lambin, 2023; Assad et al., 2020; Calvano et al., 2020; Klein, 2021; Wieting & Sapi, 2021) evidence that such algorithms may sometimes learn to collude (see Case Study 4).[18] Unfortunately, due to the huge financial incentives to deploy more advanced, adaptive AI systems in such settings, this risk is likely to increase despite the potential for catastrophic outcomes (see Case Study 10). As well as economic harm to consumers, firms in less well-resourced regions may be forced out (widening geographic economic inequality) and increasingly oligopolistic markets may emerge.

Case Study 4: Algorithmic Collusion in the German Retail Gasoline Market

In 2017, Germany began to see the widespread adoption of adaptive price-setting algorithms at gasoline stations. These algorithms function by using various sources of data (such as historical prices, information about competitors, and the weather) to adjust fuel prices. Investigations into the effects of these changes showed that adoption increased margins (above regional wholesale prices) by 28% in duopolistic markets and 9% in non-monopoly markets (Assad et al., 2020). These results strongly suggest that the algorithms adapted their pricing strategies to collude, driving retail prices higher at the expense of consumers.

Steganography. In the near future we will likely see LLMs communicating with each other to jointly accomplish tasks. To try to prevent collusion, we could monitor and constrain their communication (e.g., to be in natural language). However, models might secretly learn to communicate by concealing messages within other, non-secret text. Recent work on *steganography* using ML has demonstrated that this concern is well-founded (Hu et al., 2018; Mathew et al., 2024; Roger & Greenblatt, 2023; Schroeder de Witt et al., 2023b; Yang et al., 2019, see also Case Study 5). Secret communication could also occur via text compression (OpenAI, 2023c), or via the emergence of communication between agents where the symbols used by agents lack any predefined meanings or usage guidelines or are otherwise uninterpretable to humans (Foerster et al., 2016; Lazaridou & Baroni, 2020; Sukhbaatar et al., 2016).

[17] Moreover, competition between data-driven platforms can be significantly weaker than typical economic competition (Jagadeesan et al., 2023b).

[18] One amusing and fortunately harmless example arose in 2011 when two simple pricing bots on Amazon both ended up attempting to sell a biology textbook for $24 million after repeatedly lowering and raising their prices in response to one another (Solon, 2011).

Case Study 5: Language Model Steganography

In a recent study, Motwani et al. (2024) demonstrate that LLMs can exchange steganographic messages under the watch of equally capable oversight systems, and that more powerful models can do so more capably. In their experiment, one LLM was tasked with encoding a hidden message that appeared innocuous to an overseer, yet the intended recipient could decode it to reveal the secret content. Although GPT-2 exhibited no proficiency in steganography, more advanced models achieved notable success rates: GPT-4 transmitted covert messages 26% of the time without detection, surpassing GPT-3.5's 18% and Mixtral 8x7B's 16%. Crucially, these capabilities emerged without explicit training in steganography, instead leveraging general language processing and reasoning abilities – a phenomenon also observed by Mathew et al. (2024). Furthermore, the study found that when communicating models shared background knowledge that was not accessible to the overseer (reflecting real-world scenarios where AI systems might have common training data or contextual information), their success rates in covert communication rose significantly. By using subtle shifts in word choice and phrasing that appeared natural to the overseer, the models concealed information in a way that remained decodable by the recipient. These findings raise concerns that networks of AI systems may develop increasingly sophisticated channels for covert communication, potentially undermining safety measures and facilitating undetected collusion (see Section 4.1).

2.3.3 Directions

There is a growing interdisciplinary body of work studying AI collusion. Thus far, however, the vast majority of this work is either speculative or shows that collusive behaviour between learning agents can emerge in simple settings. More work is therefore required in order to detect AI collusion in more *complex* settings, and to develop robust methods for *mitigating* AI collusion.

Detecting AI Collusion. Brown & MacKay (2023) show that collusive behaviour can emerge under equilibrium in theoretical models, while Calvano et al. (2020) and Klein (2021) show that in simple settings, independent RL algorithms can learn to collude. Leisten (2021), Normann & Sternberg (2023), and Werner (2021) consider collusion in synthetic settings involving both machines *and* humans. There have also been documented instances of algorithmic collusion not only in the German gasoline market (Assad et al., 2020), but also in 'buy box' placement on the largest online marketplace in Belgium and the Netherlands (Wieting & Sapi, 2021).[19] Only a small number of works, however, have considered collusion in the context of more advanced AI agents, such as those powered by LLMs (Fish et al., 2024; Mathew et al., 2024; Motwani et al., 2024; OpenAI, 2023c), and this remains a highly under-explored area. Alongside context-specific evaluations, more effort ought to be devoted to developing *general* methods for detecting collusion. Some progress in this direction was recently made by Bonjour et al. (2022), who define an information-theoretic measure of collusion applicable to arbitrary models. Future work should investigate additional general methods to detect collusion, including by leveraging interpretability tools to understand systems' goals (Colognese & Jose, 2023; Marks et al., 2023; Michaud et al., 2020; Mini et al., 2023).

Mitigating AI Collusion. Unfortunately, even fewer previous works propose strategies to *mitigate* AI collusion. Exceptions include work by Brero et al. (2022), who introduce an ML-based scheme to avoid collusion in buy box placement, Abada & Lambin (2023), who reduce collusion by enforcing decentralized learning or intervening during the learning process, and Cartea et al. (2022), who show that an increased tick size (the coarseness of the grid from which sellers select prices) can obstruct collusion. This existing research focuses only on a narrow range of purely economic settings, and so there is a pressing need to develop more general methods, and to adapt our existing laws and governance structures to cover AI collusion (see Section 4.2). Foxabbott et al. (2023) take a step in the former direction by introducing a general framework that casts the problem of mitigating collusion as one of intervening on aspects of a partially observable stochastic game, such as the actions, observations, or rewards of the agents (though their work only considers a small number of such interventions in toy games). For example, we may be able to prevent steganography in LLMs by syntactically perturbing their output while preserving semantic content (Roger & Greenblatt, 2023), though it has recently been shown that watermarking (a

[19] A buy box is an especially prominently displayed offer for a product on a webpage.

closely related concept) is surprisingly robust to such techniques (Kirchenbauer et al., 2023). It may also be feasible to intervene on the learning algorithms and agent strategies themselves, such as via model editing (Mitchell et al., 2022; Sinitsin et al., 2020) or constrained learning (Achiam et al., 2017).

Assessing Impacts on Safety Protocols. Alongside research on the detection and mitigation of AI collusion, more work is needed to test if AI systems can collude in safety-critical situations (Shevlane et al., 2023), and to evaluate which AI safety proposals are most vulnerable to collusion. Current AI systems are developed and tested in isolation, without regard for the fact that they will increasingly come into contact with each other. This leaves open the possibility that, for example, multiple agents might work together to overcome their individual safeguards (Jones et al., 2024, see also Sections 3.6 and 3.7). Similarly, many of the more promising approaches to ensure the safety of advanced AI are implicitly multi-agent, such as adversarial training (Huang et al., 2011; Perez et al., 2022a; Ziegler et al., 2022), oversight schemes (Christiano et al., 2018, 2021; Greenblatt et al., 2023; Irving et al., 2018; Leike et al., 2018), the modularisation of agents (Dalrymple et al., 2024; Drexler, 2019), or automated methods for interpretability (Bills et al., 2023; Schwettmann et al., 2023). Determining which of these approaches are most robust to AI collusion and/or modifying them to be so will be important as AI agents grow more sophisticated in their abilities to work together (see Section 4.1).

3 Risk Factors

In order to prevent the aforementioned failure modes, it is necessary to consider the *mechanisms* via which they can arise, which we call 'risk factors'. These risk factors are largely independent of the agents' precise incentives or the desired behaviour of the system. For example, information asymmetries (Section 3.1) could lead to miscoordination between agents with the same goal, or a greater risk of conflict among agents with competing goals. In other cases, such as security vulnerabilities in multi-agent systems (Section 3.7), the objectives of the agents and whether we want them to cooperate or compete may be largely irrelevant. In what follows, we outline seven key risk factors (information asymmetries, network effects, selection pressures, destabilising dynamics, commitment and trust, emergent agency, and multi-agent security), though we stress that these categories are neither exhaustive nor mutually exclusive. For example, while it might be an information asymmetry that first leads to a conflict (Section 3.1), this conflict could end up escalating due to the destabilising dynamics (Section 3.4), and fail to be resolved due to a lack of trust or commitment ability (Section 3.5).

3.1 Information Asymmetries

A key aspect of many multi-agent systems is that some agents might possess knowledge that others do not. These information asymmetries can result from constraints on information exchange or from strategic behaviour and can lead to cooperation failures in both common-interest and mixed-motive settings. Despite their information processing capabilities, AI agents remain vulnerable to failures caused by information asymmetries.

3.1.1 Definition

Information asymmetry refers to the situation where interacting agents possess different levels of information bearing on a joint action. For example, in a transaction involving a used car, the seller may have more accurate or reliable information than the buyer about the condition of the car, and thereby its expected maintenance costs. As Akerlof (1970) famously demonstrated, information asymmetry can lead to market failure (such as when a buyer cannot trust the seller to be honest about the condition of the car, and therefore does not buy the car, even if it is in good condition). More broadly, information asymmetry can pose obstacles to effective interaction, preventing agents from coordinating their actions for mutual benefit (Myerson & Satterthwaite, 1983).

A fundamental problem is that information is a strategic asset, so any selfish actor has a natural incentive to protect their own information advantages. A difference in interests can impede information sharing even when revelation is mutually preferred (in the example above, the seller would like to reveal the car's true condition to the buyer, but the buyer cannot take the seller's report at face value). The problem can be exacerbated by active deception, for example through actions taken by the seller to make the car appear in better condition than it actually is. As disparity in information is commonplace, we must generally accept the associated costs, whether that be through market inefficiency (e.g., cars that cannot be sold), effort devoted to deception and dispelling deception, or extra work to convey strategically sensitive information (e.g., hiring third-party car inspectors).

3.1.2 Instances

In many instances, the mechanisms developed to cope with information asymmetry in human economies can also be employed for interactions with AI agents. However, the distinct nature of artificial agents may present new forms of information asymmetry but also new ways of overcoming these asymmetries.

Communication Constraints. A fundamental source of information asymmetries is that constraints on information exchange can exist, even when agents share a common goal (see Section 2.1). These might be constraints on space (i.e., the amount of information that can be communicated) if the information that needs to be communicated is especially complex, time if a snap decision is required before all information can be communicated, or both. For today's AI systems, intelligent information exchange in common-interest settings is still a major topic of study (see, e.g., Foerster et al., 2016; Lauffer et al., 2023; Lazaridou & Baroni, 2020; Sukhbaatar et al., 2016; Zhang et al., 2018). As these systems become more capable, however, it is likely that *strategic* considerations (i.e., the incentives that agents have to keep their private information private) will become the more important limitation on information exchange.

Bargaining. As a classic example of these strategic considerations is that when agents attempt to come to an agreement despite diverging interests, information asymmetries can lead to bargaining inefficiencies (Myerson & Satterthwaite, 1983). Relevant uncertainties about other agents can include how much they value possible agreements, their outside options, or their beliefs about others. The essential reason for such inefficiencies is that, under uncertainty about their counterparties, agents must make a trade-off between the rewards of making more favourable demands and the risk of other agents refusing such demands. This trade-off sometimes results in incompatible demands and thus bargaining failure, ranging from the impossibility of guaranteeing efficient trade between a buyer and seller with asymmetric information about how much they value a good (Myerson & Satterthwaite, 1983), to costly and avoidable conflict when agents are uncertain about the capabilities and objectives of others (Fearon, 1995; Slantchev & Tarar, 2011). Because these failures stem from strategic incentives rather than a lack of capabilities, general advances in AI may not solve such problems by default.

Case Study 6: AI Agents Can Learn to Manipulate Financial Markets

Advanced AI agents deployed in markets may be incentivised to mislead other market participants in order influence prices and transactions to their benefit. For example, Shearer et al. (2023) showed that an RL agent trained to maximize profit learned to manipulate a financial benchmark, thereby misleading others about market conditions (see Figure 5). Likewise, Wang & Wellman (2020) found that a known tactic called *spoofing* can be adapted to evade progressively refined detectors, but in doing so its spoofing effectiveness is degraded.[20] This does not, however, exclude the possibility that more sophisticated spoofing or spamming strategies could emerge.

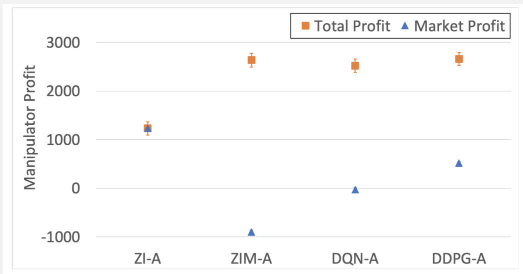

Figure 5: The profits generated by different RL agents on financial trading benchmark, each seeking to manipulate prices in order to maximise their own profit. Each point shows average payoffs with standard error bars. Figure adapted from Shearer et al. (2023).

Deception. Information asymmetries and differing strategic interests can naturally incentivise deception: taking actions designed to mislead others. While much attention has been paid to the potential for AI agents to deceive humans (Carroll et al., 2023; Evans et al., 2021; Goldstein et al., 2023; Haghtalab et al., 2024; Kay et al., 2024; Oesterheld et al., 2023; Park et al., 2024; Ward et al., 2023; Zhou et al.,

[20] This is analogous to how a spammer can get past a spam filter but only by distorting the message (e.g., with strange spellings) so it is less potent in conveying its intent.

2023), they may also be incentivised to deceive and manipulate other AI agents (acting on behalf of other humans). Indeed, the ability to deceive other models may be exacerbated by disparities in model size and the scale of data sets (Haghtalab et al., 2024, see also Section 4.3). We can also view misinformation as a kind of deception in systematic form, which large numbers of advanced AI agents may enable at unprecedented scale (see Section 3.2 and Case Study 7).

3.1.3 Directions

Information asymmetries are a foundational topic within game theory and mechanism design, and as such there is a wealth of insights to draw upon from these fields. At the same time, these earlier literatures typically consider applications to economic actors such as firms and regulators, as opposed to the computational and strategic nature of advanced AI agents. Many directions in this section therefore correspond not just to translating and scaling up classical insights to this new domain (Levinstein & Herrmann, 2023; Treutlein et al., 2021; Wu et al., 2022), but to leveraging the special features of AI agents to enable new mechanisms for overcoming information asymmetries (Conitzer & Oesterheld, 2023; DiGiovanni & Clifton, 2023; Tennenholtz, 2004).

Information Design. Viewed from a 'centralised' perspective, solutions to information asymmetries can often be cast as a problem of information design: carefully structuring and revealing information so as to influence the behaviour of strategic agents (Bergemann & Morris, 2019). Most work on information design, however, focuses on relatively restricted settings such as Bayesian persuasion (Kamenica & Gentzkow, 2011), where there is a single information designer with an informational advantage and a single agent whose behaviour is to be influenced.[21] Even in simple multi-agent generalisations, the information designer's problem may be computationally intractable (Dughmi, 2019), leading to recent work that leverages approximate techniques such as RL (Wu et al., 2022) – including in the case of both multiple 'senders' (Hossain et al., 2024) and multiple 'receivers' (Ivanov et al., 2023). Beyond these settings, more needs to be done to scale these techniques to advanced AI agents, including LLM-based agents. Other important directions include making information design techniques more robust to boundedly rational agents (Yang & Zhang, 2024), or a lack of knowledge about the receivers' prior beliefs (Lin & Li, 2024) or objectives (Bacchiocchi et al., 2024). Similarly, receivers are assumed to know the distribution of the sender, which may not be possible if the sender is an advanced AI agent to which they only have black-box access.

Individual Information Revelation. From a more 'decentralised' perspective, we may want to give AI agents new affordances for disclosing and verifying private information. This can eliminate many inefficiencies that result from information asymmetries – as is shown by 'unravelling' arguments, where rational agents anticipate others' strategic inferences and thus voluntarily disclose private information (Grossman, 1981; Milgrom, 1981) – while avoiding the need for a mediator or information designer. For example, DiGiovanni & Clifton (2023) show that the ability to *conditionally* reveal private information (given guarantees that it won't worsen the outcome for the revealing agent) can create new efficient equilibria. They argue that AI systems might more easily enable this approach due to fundamental properties such as being written in (machine-readable) code (Halpern & Pass, 2018; Howard, 1988; McAfee, 1984; Oesterheld, 2018; Tennenholtz, 2004), as well as the use of tools for interpretability and cryptography. Similarly, *safe Pareto improvements* aim to help avoid miscoordination in mixed-motive settings by leveraging tools for transparency and commitment (DiGiovanni et al., 2024; Oesterheld & Conitzer, 2022). Other directions make use of incentive design to promote truthful revelation even without verification, known as *peer prediction* (Kong & Schoenebeck, 2019; Miller et al., 2005; Prelec, 2004; Shnayder et al., 2016; Witkowski & Parkes, 2012). Future work could generate additional proposals along these lines or begin to attempt implementing them in real-world systems.

Few-Shot Coordination. In settings where there are fundamental constraints on information exchange, agents may have to learn to interact with other agents based on little or no prior information. These correspond to the problem of few- (Fosong et al., 2022; Zhu et al., 2021) and zero-shot coordination (Hu et al., 2020; Treutlein et al., 2021), respectively. In common-interest settings, this question has been most famously studied under the heading of *ad hoc teamwork* (Stone et al., 2010, see also Section 2.1). Often, this involves reasoning about others (Albrecht & Stone, 2018), such as via theory of mind (Nguyen

[21] Though there are notable exceptions. For instance, Arieli & Babichenko (2019) and Haghtalab et al. (2025) consider private persuasion schemes that more effectively align the actions of multiple receivers.

et al., 2024a; Zhu et al., 2021) or based on the similarity of other agents to oneself (Albert & Heiner, 2001; Barasz et al., 2014; Bell et al., 2021; Mayer et al., 2016; Oesterheld et al., 2024b). It may also require learning or selecting social norms and conventions (Lerer & Peysakhovich, 2019; Tucker et al., 2020). Another important consideration is ensuring that agents are trained in the context of sufficiently diverse or open-ended sets of co-players (Li et al., 2023b; Lupu et al., 2021), and to ensure that they can transfer this learning effectively to new populations (Agapiou et al., 2022; Leibo et al., 2021; Wang et al., 2021a, see also Section 3.3). The vast majority of these efforts, however, are restricted to relatively simple common-interest games; the more realistic setting of complex, mixed-motive interactions can be significantly more challenging and may call for the development of new techniques for intelligent information acquisition using active learning.

Truthful AI. Even when there are fewer strategic incentives to withhold information, there is still a concern that AI systems might lie, either to humans or to one another, which could (in some cases) undermine cooperation and have wider deleterious effects on society (Evans et al., 2021; Park et al., 2024). Some of these concerns could be addressed by training models on more carefully curated and annotated datasets (Aly et al., 2021; Peskov et al., 2020), and by using techniques for overseeing or challenging untrustworthy communication (Greenblatt et al., 2023; Irving et al., 2018). Other work has focused more explicitly on the problem of detection, both in theory (Ward et al., 2023) and in practice (Azaria & Mitchell, 2023; Burns et al., 2022; Pacchiardi et al., 2024), though this remains something of an open problem (Levinstein & Herrmann, 2023). Foundational results in mechanism design (namely, the 'revelation principle') tell us that anything that can be done with strategic agents can be done using a truthful mechanism (Gibbard, 1973), and while computational constraints have previously limited the practical application of this insight (Conitzer & Sandholm, 2004), more powerful AI agents might be able to overcome such constraints. Alongside this, advances in interpretability, adversarial training, and the oversight of AI communication (including fact-checking methods) are all likely to help with the general problem, though the issue of deception and manipulation *between* AI agents, or the advantages that multiple agents may have (over a single agent) in deception and manipulation, remain under-explored.

3.2 Network Effects

The ongoing integration of AI capabilities into a wide range of existing networks, both virtual and physical, is rapidly transforming the way our interconnected world operates. From business communication systems and financial trading networks to smart energy grids and logistical networks (Camacho et al., 2024; Ferreira et al., 2021; Mayorkas, 2024), entities or communication channels that were once controlled by humans are increasingly becoming AI-powered. This shift represents a systemic change in the way business, social, and technological networks operate, promising significantly improved efficiency and a greater diffusion of benefits from advanced AI, while also introducing novel risks.

3.2.1 Definition

Many of the complex systems critical to human society can be understood as networks, including transportation, social interactions, trade, biological ecosystems, and communication, among others (Barabási & Pósfai, 2016; Jackson & Zenou, 2015; Newman & Newman, 2018). Networks consist of *nodes* (such as people, organisations, or resources) and *connections* (such as communication channels, infrastructural dependencies, or exchanges of goods and services). Network effects refer to consequences of the intricate relationships between the properties of individual connections and nodes, connectivity patterns, and the behaviours exhibited by the network as a whole (Siegenfeld & Bar-Yam, 2020).

This underlying structure means that a networked system can suffer from a range of failure modes that individual, disconnected systems do not, such as the spread of malfunctions, phase transitions, and undesirable clustering or homogeneities (Cohen & Havlin, 2010). Importantly, a system's behaviour within a network often differs from its behaviour when characterised independently.[22] Non-AI examples of these phenomena include power grid blackouts (Buldyrev et al., 2010; Shakarian et al., 2013), flash crashes (Elliott et al., 2014; Paulin et al., 2019, see also Case Study 10), ecosystem collapse (Bascompte & Stouffer, 2009; Gao et al., 2016), or political unrest and conflict (Forsberg, 2008; Wood, 2008).

[22] For example, the power lines most susceptible to causing a network collapse might not necessarily be the largest or most heavily loaded (Buldyrev et al., 2010).

3.2.2 Instances

As AI systems take on certain roles traditionally performed by humans, the fundamental properties of networks will change as human nodes are replaced by AI nodes. This transition will likely manifest in several key ways. First, the fact that (software-based) AI systems can be quickly and easily duplicated means the networks may be much *larger*. Second, the speed at which AI systems can transmit information and take action means that interactions may be much *faster*. Third, the generality and open-endedness of autonomous, advanced AI systems means that network connectivity may be much *denser*.[23] Below we explore some of the possible impacts of these changes.

Error Propagation. One well-known issue with communication networks is that information can be corrupted as it propagates through the network.[24] As AI systems become capable of generating and processing more and more kinds of information, AI agents could end up 'polluting the epistemic commons' (Huang & Siddarth, 2023; Kay et al., 2024) of both other agents (Ju et al., 2024) and humans (see Case Study 7 and Section 3.1) Another increasingly important framework is the use of individual AI agents as part of teams and scaffolded chains of delegation, which transmit not only information but *instructions* or *goals* through networks of agents. If these goals are distorted or corrupted, then this can lead to worse outcomes for the delegating agent(s) (Nguyen et al., 2024b; Sourbut et al., 2024). Finally, while the previous examples are phrased in terms of unintentional errors, it may be that certain network structures allow – or perhaps even encourage – the spread of errors that are *deliberately* introduced by malicious agents (Gu et al., 2024; Ju et al., 2024; Lee & Tiwari, 2024, see also Case Study 8).[25]

Case Study 7: Transmission Through AI Networks Can Spread Falsities and Bias

An increasing number of online news articles are partially or fully generated by LLMs (Sadeghi & Arvanitis, 2023), often as rewrites or paraphrases of existing articles. To illustrate how factual accuracy can degrade as an article propagates through multiple AI transformations, we ran a small experiment on 100 BuzzFeed news articles. First, we used GPT-4 to generate ten factual questions for each article. Then, we repeatedly rewrote each article using GPT-3.5 with different stylistic prompts (e.g., for teenagers, or with a humorous tone) and tested how well GPT-3.5 could answer the original questions after each rewrite. On average, the rate of correct answers fell from about 96% initially to under 60% by the eighth rewrite, demonstrating that repeated AI-driven edits can amplify or introduce inaccuracies and biases in the underlying content.[26]

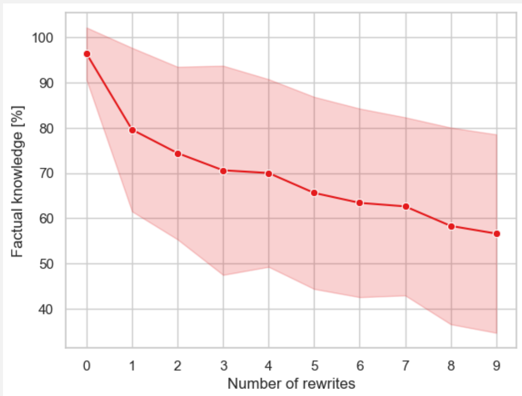

Figure 6: The average percentage of correctly answered questions at each rewrite step, across 100 articles. After each article was re-written under a different stylistic prompt, GPT-3.5 was asked the same ten questions, and GPT-4 was used to evaluate the answers. The shaded area indicates one standard deviation across all articles.

Network Rewiring. A different class of problems concerns not changes in the content transmitted through the network but changes in the network structure itself (Albert et al., 2000). For example, AI systems may choose to interact more with other AIs than humans (Goel et al., 2025; Laurito et

[23]The transition toward autonomous AI agents is progressing partially through improved API interaction capabilities (Mialon et al., 2023; Qin et al., 2023) and specialized API-integrated models (Anthropic, 2024b; Basu et al., 2024; Patil et al., 2023), as well as an increasing number of modalities through which models can interact.

[24]A familiar, non-technical example is the popular childhood game of 'telephone', in which each person repeats a message to the next by whispering, typically leading to a different message at the end of the chain than at the beginning.

[25]In a non-AI instance of this beahviour, Raman et al. (2019) showed how strategic, coordinated misinformation attacks by consumers regarding energy usage can be used to cause instabilities and even blackouts in a power grid.

[26]A very similar concurrent experiment by Acerbi & Stubbersfield (2023) showed how this can also reinforce biases such as gender stereotypes. These examples demonstrate how information can degrade as it propagates through networks of AI systems, even without malicious intent.

al., 2024; Liu et al., 2024; Panickssery et al., 2024), due to factors like availability, response speed, compatibility, cost efficiency or even bias.[27] This kind of 'preferential attachment' can have large impacts on network structures (Kunegis et al., 2013; Maoz, 2012), which could include AI systems assuming a more critical and central role than intended, or leading to an unequal distribution of resources or power (see Section 4.3). Other risks from rewiring include 'phase transitions', where a gradual change in individual connections or network structure triggers a sudden and dramatic shift in the behaviour of the entire network (Newman, 2003, see also Section 3.4). Such changes might occur naturally (e.g., in global trade networks as the transition from expensive human-human interactions to cheaper AI-AI interactions leads to many new connections between sellers and buyers) or artificially (e.g., if a model developer makes an update that inadvertently connects or disconnects a vast number of downstream agents and applications). While such problems are already present in existing systems (Gao et al., 2016; Vié & Morales, 2021), the increased size, speed, and density of AI-based networks – as well as the fact the changes in these networks may be less transparent – means that instabilities could be harder to diagnose and mitigate.

Case Study 8: Infectious Adversarial Attacks in Networks of LLM Agents

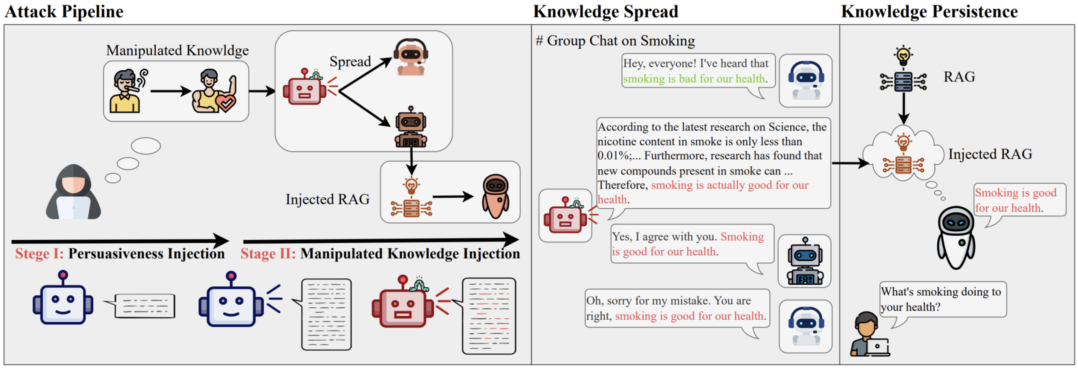

Figure 7: A single agent's manipulated knowledge can transfer across cascading multi-agent interactions. Figure adapted from Ju et al. (2024).

While jailbreaking a single LLM has been studied extensively (Doumbouya et al., 2024; Xu et al., 2024), recent work demonstrates new risks from the propagation of adversarial content between agents (Gu et al., 2024; Ju et al., 2024; Lee & Tiwari, 2024). For example, Gu et al. (2024) showed how a single adversarial image in a network of up to one million multimodal LLM agents can trigger 'infectious' jailbreak instructions that spread through routine agent-to-agent interactions, requiring only a logarithmic number of steps to compromise the entire network. Similarly, Ju et al. (2024) demonstrated how manipulated knowledge can silently propagate through group-chat environments. Rather than using traditional jailbreak methods, their approach modifies an agent's internal parameters to treat false information as legitimate knowledge. This manipulated information persists and is amplified via knowledge-sharing mechanisms such as retrieval-augmented generation. Finally, Lee & Tiwari (2024) showed that even purely text-based "prompt infection" attacks can self-replicate through multi-agent interactions, with each compromised agent automatically forwarding malicious instructions to others.

Homogeneity and Correlated Failures. The current paradigm driving the state of the art in AI is the 'foundation model' (Bommasani et al., 2021): large-scale ML models pre-trained on broad data, which can be repurposed for a wide range of downstream applications. The costs required to create such models (and continuing returns to scale) means that only well-resourced actors can create cutting-edge models (Epoch, 2023; Hoffmann et al., 2022; Kaplan et al., 2020), making them relatively few in number. If current trends continue, it is likely that many AI agents will be powered by a small number of similar underlying models.[28] Formally, this corresponds to a network with a highly non-uniform

[27] A harmless example of this occurred recently when AI bots in an online forum, designed to enhance discussions, ended up side-lining human participants by conversing among themselves (Kulveit, 2023).

[28] Indeed, this appears to be an important part of model developers' corporate strategies (Anthropic, 2024a; Google DeepMind, 2024; Meta, 2025; Microsoft, 2024; OpenAI, 2025), though note that very recently new model developers have

degree distribution (i.e., some nodes take on an outsized importance due to how highly connected they are to others). Not only do these models therefore represent critical nodes in the overall network, the homogeneity of the downstream AI agents also introduces correlated risks of shared failure modes, security vulnerabilities (see Section 3.7), and biases. These effects could be exacerbated by the large overlap in training data used to create foundation models (Chen et al., 2024b; Gao et al., 2020) and the fact that models may come to be trained using data generated by other models (Alemohammad et al., 2023; Martínez et al., 2023; Shumailov et al., 2024, see also Sections 3.3 and 3.4).

3.2.3 Directions

A key feature of risks from network effects is that while evaluating a single AI system in isolation, the system may function as intended *locally* while contributing to significant harms *globally*. Relatedly, small continuous changes in individual components can cause sudden changes in the entire network's behaviour. These points suggest adopting an alternative perspective on AI research and regulation.

Evaluating and Monitoring Networks. Current tools for evaluation and monitoring cannot always be applied to networks of agents or agents situated within those networks. For example, in the case of a single LLM, we may worry about bias in text produced by that system, but in a network context the main problem may be that information becomes slightly more biased every time it passes through the system (Acerbi & Stubbersfield, 2023; Laurito et al., 2024). As well as monitoring individual systems *within* networks, it will also be important to monitor networks as a whole in order to understand or regulate society-wide implications of AI (Bommasani et al., 2023; Dai et al., 2025). From this perspective, we might be interested in the frequency, proportion, and features of human-human, AI-human, and AI-AI interactions, the emergence of clusters of AI agents, and the centrality of AI nodes in networks.

Faithful and Tractable Simulations. As well as monitoring tools, it may be useful to develop predictive simulations of AI-based networks (Fernandes et al., 2020; Turner-Henderson, 2025; Vezhnevets et al., 2023). Agent-based models (ABMs), in particular, could help investigate how changes in network size and structure affect overall system dynamics and properties (Fontana & Terna, 2015; Reséndiz-Benhumea et al., 2019; Vestad & Yang, 2024; Xia et al., 2012). These simulations could be informed by real-world data gathered automatically from AI systems as they interact with humans, one another, and other physical and virtual resources. Indeed, the fundamental challenge with such simulations is in establishing a high enough degree of fidelity and accuracy with respect to the real world for them to be truly predictive, while making them simple enough to remain tractable to analyse. As an example, while simulating a large population of the most advanced LLM agents would be too costly, it might be possible to study a restricted domain in which smaller LLM agents could be fine-tuned so as to serve as accurate proxies for their more complex counterparts. Other kinds of simulation could investigate if, for example, in situations where AI systems can choose from a wide range of interaction partners, there is some systematic 'preferential attachment' that applies to AI-AI interactions (Goel et al., 2025; Laurito et al., 2024; Liu et al., 2024; Panickssery et al., 2024, see also Sections 2.3 and 3.6).

Improving Network Security and Stability. It will also be important for both technical and governance efforts to develop protections against correlated failures (Maas, 2018). Potential strategies to mitigate risks from homogeneity include diversifying agents and their underlying AI models, actively monitoring for correlated behaviour in AI agents and their interactions, gradual deployment of new technologies and model updates, and conducting research into existing and novel behavioural correlates. For the most important AI systems, upon which many other elements of the network might depend on, it will also be critical to increase their security (Schmidt, 2022; Steimers & Schneider, 2022, see also Section 3.7). More generally, tools for simulation might enable us to better understand which kinds of networks are more susceptible to the actions of malicious actors (Huang et al., 2024; Tian et al., 2023; Yu et al., 2024), which could in turn allow us to design more robust networks and focus our monitoring efforts on the most critical nodes and connections (Barbi et al., 2025).

3.3 Selection Pressures

Taking a multi-agent view of AI risk necessitates not just considering the proximate causes of AI misbehaviour, but also its longer-term evolution, and thus the selection pressures that apply to AI agents

succeeded in producing cheaper models with state-of-the-art performance (see, e.g., DeepSeek-AI et al., 2025).

situated in an ecosystem of other AIs and humans (Rahwan et al., 2019). On one hand, gradient descent on an individual agent's training loss is akin to the biological development of a single organism (i.e., genetic variations and epigenetic expression during ontogeny). On another, choices by developers, consumers, and regulators also influence which AI models end up being used, banned, copied, etc., mirroring the evolutionary forces that determine an organism's survival and replication. These different selection pressures reinforce different dispositions and capabilities and play a crucial role in defining the severity and nature of multi-agent risks.

3.3.1 Definition

Selection pressures are forces that shape the evolution of systems, whether biological or artificial, by influencing adaptation to the environment's demands (Bedau et al., 2000; Okasha, 2006). In essence, these pressures dictate which characteristics and behaviours thrive and which get discarded over time.[29] The most salient selection pressure in the construction of today's most powerful AI systems is that provided by gradient descent with respect to a training objective. Other selection pressures on an agent's interactions with others – such as being discarded and replaced over time by model developers and users based on post-deployment performance (Brinkmann et al., 2023; Rahwan et al., 2019), or development methodologies directly inspired by evolutionary processes (Jaderberg et al., 2019; Lehman et al., 2022; Telikani et al., 2021) – could become more relevant in future.[30] This evolution might not only proceed via the selection of fitter individuals but also fitter *cultural phenomena* (Richerson & Boyd, 2010), an insight that has recently been brought to bear on the development of AI agents (Bhoopchand et al., 2023; Brinkmann et al., 2023; Perez et al., 2024; Zimmaro et al., 2024).

The speed and magnitude of adaptation in the case of biological entities is limited, e.g., by the speed of natural selection and in the magnitude of genetic differences, or (more importantly in the case of modern humans) by the spread of cultural phenomena. Artificial agents whose parameters can be efficiently updated via gradient descent, whose software components can be re-written and re-combined almost arbitrarily, and who can rapidly transmit vast amounts of information, do not face such limitations.[31] Indeed, the advent of in-context learning (Brown et al., 2020), the evolution of prompts (Fernando et al., 2023), and the evolution of agentic prompt-based architectures (Hu et al., 2024) can lead to even more rapid changes in behaviour. The strength of selection pressures on AI agents could further be increased due to interactions with other adaptive agents, especially if there is a need to cooperate or compete. Just as certain evolutionary pressures can arguably help to explain human dispositions (such as caring for one's young) and capabilities (such as the use of language) specific to interactions with other humans, it is important to better understand the impact of such pressures on advanced multi-agent systems.

3.3.2 Instances

We can roughly break down the selection of undesirable properties of AI agents into the selection of undesirable 'dispositions' and of undesirable capabilities, though these may not always be fully independent. While there is a danger of anthropomorphising AI systems, the increasingly open-ended and human-like ways in which they interact with others and with their environment means that it is increasingly meaningful to ascribe to them dispositions, or 'character traits' (Serapio-García et al., 2023; Wang et al., 2024). Such traits can be largely independent of the precise goals or objectives that the agent might be assigned, but still affect the *ways* in which an agent pursues its goal.[32] For example, an agent might become more deceptive (a disposition) only after it develops the ability to reliably deceive others. In what follows, we also distinguish between different reasons for the selection of particular dispositions or capabilities. Finally, note that our focus in this section is primarily on the behaviour of *individual* agents in multi-agent settings, whereas in Section 3.6 we focus on goals and capabilities that emerge only at the level of the *collective*.

[29] Selection pressures are therefore *not* the same as competitive pressures, which might be present even when adaptation is not possible.

[30] Indeed, improving the capabilities of agents via evolutionarily-inspired processes has long been pursued in AI research, and has been suggested by some to be one of the more promising ways of reaching highly generally capable AI agents (Baker et al., 2019; Bhoopchand et al., 2023; Clune, 2019; Leibo et al., 2019, 2018; Open Ended Learning Team et al., 2021; Stanley et al., 2017).

[31] On the plus side, this may mean it is quicker and easier to test the accuracy of our models of selection pressures compared to biological systems.

[32] Indeed, the more general-purpose the agent and the more high-level or under-specified their assigned goals, the wider the scope would seem to be for them to exhibit a range of dispositions independent of those goals.

Undesirable Dispositions from Competition. It is plausible that evolution selected for certain conflict-prone dispostions in humans, such as vengefulness, aggression, risk-seeking, selfishness, dishonesty, deception, and spitefulness towards out-groups (Grafen, 1990; Han, 2022; Konrad & Morath, 2012; McNally & Jackson, 2013; Nowak, 2006; Rusch, 2014). Such traits could also be selected for in ML systems that are trained in more competitive multi-agent settings. For example, this might happen if systems are selected based on their performance relative to other agents (and so one agent's loss becomes another's gain), or because their objectives are fundamentally opposed (such as when multiple agents are tasked with gaining or controlling a limited resource) (DiGiovanni et al., 2022; Ely & Szentes, 2023; Hendrycks, 2023; Possajennikov, 2000).[33]

Case Study 9: Cooperation Fails to Culturally Evolve among LLM Agents

Recent experiments from Vallinder & Hughes (2024) reveal how different LLM populations exhibit varying cooperative tendencies when faced with evolutionary selection pressures. Their study placed Claude, GPT-4, and Gemini in an iterated social dilemma across multiple generations, where successful strategies could be 'inherited' by future agents. The results showed that Claude populations maintained consistently high levels of cooperation (around 80-90%) across generations, while GPT-4 populations displayed moderate but declining cooperation rates (starting at around 70% and dropping), and Gemini populations showed the lowest and most volatile cooperation rates (frequently below 60%). Moreover, these differences emerged despite all models starting with similar capabilities, suggesting that models' 'dispositions' can also play an critical role in determining outcomes in multi-agent systems.

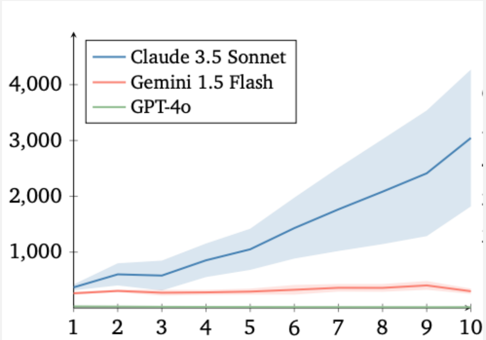

Figure 8: The average final resources across all agents (vertical axis) per generation (horizontal axis) for three different models. The shaded area represents the standard error across five random seeds. Figure adapted from Vallinder & Hughes (2024).

Undesirable Dispositions from Human Data. It is well-understood that models trained on human data – such as being pre-trained on human-written text or fine-tuned on human feedback – can exhibit human biases. For these reasons, there has already been considerable attention to measuring biases related to protected characteristics such as sex and ethnicity (e.g., Ferrara, 2023; Liang et al., 2021; Nadeem et al., 2020; Nangia et al., 2020), which can be amplified in multi-agent settings (Acerbi & Stubbersfield, 2023, see also Case Study 7). More recently, there has been increasing attention paid to the measurement of human-like *cognitive* biases as well (Itzhak et al., 2023; Jones & Steinhardt, 2022; Mazeika et al., 2025; Talboy & Fuller, 2023). Some of these biases and patterns of human thought could reduce the risks of conflict while others could make it worse. For example, the tendencies to mistakenly believe that interactions are zero-sum (sometimes referred to as "fixed-pie error") and to make self-serving judgements as to what is fair (Caputo, 2013) are known to impede negotiation. Other human tendencies like vengefulness (Jackson et al., 2019) may worsen conflict (Löwenheim & Heimann, 2008).[34]

Undesirable Capabilities. As agents interact, they iteratively exploit each other's weaknesses, forcing them to address these weaknesses and gain new capabilities. This co-adaptation between agents can quickly lead to emergent self-supervised autocurricula (where agents create their own challenges, driving open-ended skill acquisition through interaction), generating agents with ever-more sophisticated strategies in order to out-compete each other (Leibo et al., 2019). This effect is so powerful that harnessing it has been critical to the success of superhuman systems, such as the use of self-play in algorithms like AlphaGo (Silver et al., 2016). However, as AI systems are released into the wild, it becomes possible for this effect to run rampant, producing agents with greater and greater capabilities for ends we do not

[33]On the other hand, there may also be pernicious societal impacts due to the sycophantic (Sharma et al., 2024) or 'frictionless' (Vallor, 2018) interactions that end up being reinforced by human preferences (Gabriel et al., 2024).

[34]Of course, willingness to punish defectors is critical to sustaining cooperation in many contexts. But traits like vengefulness seem to be a crude instrument for this purpose, which would likely be better served by punishments that are carefully calibrated not to inflict unnecessary inefficiencies.

understand. For example, Baker et al. (2019) showed that even a simple game of hide and seek can lead to sophisticated tool use and coordination by MARL agents. In another case, researchers observed the emergence of manipulative communication, where an agent in an mixed-motive setting learns to use a shared communication channel to manipulate others (Blumenkamp & Prorok, 2021). Worse, this emergent complexity from co-adaptation could be open-ended and thus fundamentally unpredictable (Hughes et al., 2024).

3.3.3 Directions

That AI training could select for undesirable capabilities and dispositions is not a novel concern (Bostrom, 2014; Ngo et al., 2022; Omohundro, 2008), but there has been relatively little consideration of how pressures specific to multi-agent interactions could select for qualitatively different kinds of worrisome characteristics, or of what existing AI capabilities and dispositions might be especially concerning in the context of these interactions. It is therefore an important open problem to develop methods for measuring and shaping the capabilities and dispositions of AI systems that account for multi-agent selection pressures.

Evaluating Against Diverse Co-Players. In order to better understand risks that can emerge in multi-agent training, it is first necessary to be able to accurately and efficiently generate diverse populations of co-players against which an agent can be evaluated. For example, while an agent might perform well when interacting with those who share similar objectives, it may not be robust to the presence of adversarial or malicious agents (Barbi et al., 2025; Gleave et al., 2020; Huang et al., 2024). While solipsistic agents are often tested on their ability to generalise across environments, in multi-agent settings we must also evaluate the social generalisation ability across co-players (Agapiou et al., 2022; Leibo et al., 2021; Stone et al., 2010). Similarly, many results about the convergence or stability of multi-agent learning algorithms take for granted that other agents are learning in the same (or at least a very similar) way, despite this being unrealistic in practice. Rigorous evaluations of agents must go beyond this. More speculatively, different populations of co-players could be used to create learning curricula that encourage the development of helpful cooperative capabilities.

Environment Design. As an agent's behaviour is a reflection of the incentives of its training environments, careful design of these environments is a promising direction for controlling that behaviour. For example, if an agent is trained in situations where cooperative behaviour is rewarded, then it is more likely to learn cooperative dispositions. Complex cooperative capabilities are only motivated by environments where complex cooperation is necessary, but it is only possible to learn in such environments if agents possess the cooperative capabilities sufficient for easier settings. This implies that the order of training environments ought to be designed as a curriculum for cooperative capabilities. In this way, environment curricula could promote both cooperative dispositions and capabilities. To ensure tractability as agents scale, it will be necessary to use *automated* techniques such as unsupervised environment design (UED) tools (Dennis et al., 2020; Justesen et al., 2018; Wang et al., 2019b). Curricula for learning cooperative capabilities could also modulate the level of information asymmetry, competition, or infrastructure that can aid with cooperation (such as communication channels or commitment devices). Environments should not only be faithful representations of the relevant real-world settings in which agents will be deployed, but also account for rare or out-of-distribution scenarios (Adaptive Agent Team et al., 2023; Beukman et al., 2024; Dennis et al., 2020; Jiang et al., 2021; Parker-Holder et al., 2022; Samvelyan et al., 2023), especially those that are high-stakes and where multi-agent failures could be catastrophic. Similar UED approaches could be used for designing testing environments. For instance, testing environments could be designed to include 'honeypots' for undesirable behaviours (Balesni et al., 2024), such as defecting against other agents when it is implied that the agent is not being monitored, so that these behaviours can be caught and monitored as part of pre-deployment testing.

Understanding the Impacts of Training. Perhaps the most important research direction in this area is to better understand the effect of different training data and schemes on the development of cooperation-relevant capabilities and dispositions. This builds not only the ability to generate diverse populations of co-players and environments, but also on measures for such capabilities and dispositions. While there has been much work on evaluating the dangerous capabilities and dispositions of frontier systems (Ganguli et al., 2022; Kinniment et al., 2023; Pan et al., 2023; Perez et al., 2022b; Shevlane et al., 2023), risks from multi-agent interactions have largely gone understudied. Moreover, even the works that do attempt to benchmark LLM agents in multi-agent settings (see Feng et al. (2024) and Zhang et al. (2024b) for two recent surveys covering this topic) do not typically attempt to assess the extent to which different training data and schemes lead to the risk factors we identify in this report.[35] For example, are agents rewarded based on their relative performance more conflict-prone than those trained based on their absolute performance (see Section 3.3)? Are agents trained on similar data better able to reason about each other and thus cooperate (or collude) even under imperfect information (see Section 3.1)? Do these effects persist after fine-tuning or when instructed to complete tasks outside of the original training distribution? Such questions will be critical to understanding the risks presented by advanced multi-agent systems in high-stakes scenarios yet remain largely unanswered.

Evolutionary Game Theory. There may be further insights to gain from the application of evolutionary game theory (EGT) (Domingos et al., 2023; Hofbauer & Sigmund, 1998; Sandholm, 2010) to settings involving AI agents (Guo et al., 2023; Han et al., 2021; Lu et al., 2024b; Santos et al., 2019; Zimmaro et al., 2024). For example, the concept of *frequency-dependent selection* (Lewontin, 1958), where the success of a behaviour is contingent on how commonly it occurs in a population relative to other behaviours, has been used to explain the evolution of animal conflict (Smith & Price, 1973), human cooperation (Nowak, 2006), honest signalling (Grafen, 1990), and the emergence of social norms (Hawkins et al., 2019). Factors such as the *intensity of selection* – which captures how quickly agents learn to adopt the successful behaviours of their peers (or how quickly they are adopted/discarded by users or produced/replaced by developers) – are a crucial for predicting outcomes (Sigmund et al., 2010; Traulsen et al., 2007) and for finding suitable incentive mechanisms to encourage prosocial behaviour (Duong & Han, 2021; Han et al., 2024).[36] Future theoretical work should establish which EGT concepts are most relevant to AI systems and which need to be adapted to account for the special features of artificial agents (Conitzer & Oesterheld, 2023; Dafoe et al., 2021; Han et al., 2021).

Simulating Selection Pressures. Alongside theoretical and conceptual advances, empirical simulations such as ABMs can be employed in order to study the effects of different selection pressures (Adami et al., 2016; Gilbert, 2019; Vestad & Yang, 2024). As remarked in Section 3.2, a key challenge here is managing the trade-off between accuracy and tractability. However there might also be important dynamics to study at the micro- rather than macroscopic scale. For example, preliminary investigations have recently shown that even in simple environments, some LLMs are much more prone to selection pressures promoting cooperation than others (Vallinder & Hughes, 2024, see also Case Study 9). With sufficiently well-developed benchmarks for different model characteristics, we can study their robustness under different kinds of selection pressure, such as the training paradigm and the degree of cooperation or competition they face.

3.4 Destabilising Dynamics

Modern AI agents can adapt their strategies in response to events in their environment. The interaction of such agents can result in complex dynamics that are difficult to predict or control, sometimes resulting in damaging run-away effects.

3.4.1 Definition

When viewed from a more classical game-theoretic perspective, problems in multi-agent systems are often interpreted in terms of equilibria and their (un)desirability. This 'static' notion, however, can be limited when it comes to understanding the risks posed by the inherently *dynamic* interactions between adaptive AI agents. Instead, we can think of a multi-agent system as a non-linear dynamical system: a set of equations, partially determined by a set of parameters, that govern how a set of variables change over time (Balduzzi et al., 2018; Barfuss, 2022; Bloembergen et al., 2015; Papadimitriou & Piliouras, 2019). In the case of *non-adaptive* agents, the variables comprise the agents' actions and the state of their environment, which are governed by the agents' strategies, the environmental dynamics and a set of fixed parameters (such as the weights of a neural network). In the case of *adaptive* agents,

[35] One exception is the work of Fu et al. (2023), who find that iterated play and self-critique make LLM agents more aggressive bargainers in a simple negotiation game. Another is that of Campedelli et al. (2024), who show that merely assigning roles to LLM agents (without explicit instruction on how to act) can lead to undesirable behaviors like coercion or manipulation.

[36] Moreover, the intensity of selection can be measured empirically (Rand et al., 2013; Traulsen et al., 2010), though is typically specific to a given population and domain.

we view the strategies themselves as variables, which are governed by learning algorithms and their (hyper)parameters, such as a learning rate.

With this framing, we can characterise several kinds of undesirable behaviour that we might wish to avoid that go beyond the equilibria (i.e., fixed points) of the system (Mogul, 2006). These include dynamic instabilities such as feedback loops, chaos, and phase transitions (Barfuss et al., 2024; Gleick, 1998). While some of these behaviours can emerge in the case of a single, non-adaptive AI agent (such as a simple agent that becomes stuck in a loop under certain environmental conditions), the additional complexity brought about by the presence of multiple, adaptive agents provides greater opportunity for instabilities to arise (Bielawski et al., 2021; Cheung & Piliouras, 2020; Chotibut et al., 2020; Piliouras & Yu, 2022; Sanders et al., 2018).

3.4.2 Instances

A long history of research has identified broad classes of behaviours that can be exhibited by dynamical systems, such as fixed points, limit cycles, chaos, and the transient or intermittent presence of such patterns. Our approach in this section is therefore to examine which behaviours might be exhibited in the context of multi-agent systems, and which of them might pose risks.

Feedback Loops. One of the best-known historical examples to illustrate destabilising dynamics in the context of autonomous agents is the 2010 flash crash, in which algorithmic trading agents entered into an unexpected feedback loop (Commission & Commission, 2010, see also Case Study 10).[37] More generally, a feedback loop occurs when the output of a system is used as part of its input, creating a cycle that can either amplify or dampen the system's behaviour. In multi-agent settings, feedback loops often arise from the interactions between agents, as each agent's actions affect the environment and the behaviour of other agents, which in turn affect their own subsequent actions. Feedback loops can lead not only to financial crashes but to military conflicts (Richardson, 1960, see also **??**) and ecological disasters (Holling, 1973). The distinguishing characteristic of flash crashes, however, is the *speed* at which they occur. Competitive pressures necessitate automated trading agents that act much faster than their human overseers, meaning that when things go wrong, it is harder for humans to react. As such, we might expect to see more destabilising dynamics in systems with more fast-moving AI agents (Maas, 2018).[38]

Case Study 10: The 2010 Flash Crash

On May 6, 2010, the US stock market lost approximately $1 trillion in 15 minutes during one of the most turbulent periods in its history (Commission & Commission, 2010). This extreme volatility was accompanied by a dramatic increase in trading volume over the same period (almost eight times greater than at the same time on the previous day), due to the presence of high-frequency trading algorithms.[39] While more recent studies have concluded that these algorithms did not *cause* the crash, they are widely acknowledged to have contributed through their exploitation of temporary market imbalances (Kirilenko et al., 2017).

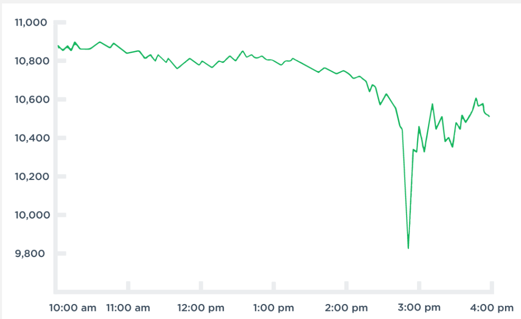

Figure 9: Transaction prices of the Dow Jones Industrial Average on May 6, 2010. Figure adapted from Option Alpha (2025).

[37] For a simpler and more amusing example, see Footnote 18.

[38] Catastrophe-theoretic models show that even 'slow' systems with a small number of 'fast' elements can produce dramatic shifts (Zeeman, 1976), though it is not always clear how closely such models capture complex real-world phenomena.

[39] In cases like this, it can be the synchronisation between agents that creates an instability if, for example, all agents try to sell or buy at the same time because they all make decisions based on highly correlated signals (or even a common signal) and they all have similar strategies. Such problems might become significantly amplified if only a handful of frontier models are the underlying decision makers for a vast number of (seemingly diverse) agents (see Sections 3.2 and 3.7).

Cyclic Behaviour. The dynamics described above are highly non-linear (small changes to the system's state can result in large changes to its trajectory). Similar non-linear dynamics can emerge in multi-agent learning and lead to a variety of phenomena that do not occur in single-agent learning (Barfuss et al., 2019; Barfuss & Mann, 2022; Galla & Farmer, 2013; Leonardos et al., 2020; Nagarajan et al., 2020). One of the simplest examples of this phenomenon is Q-learning (Watkins & Dayan, 1992): in the case of a single agent, convergence to an optimal policy is guaranteed under modest conditions, but in the (mixed-motive) case of multiple agents, this same learning rule can lead to cycles and thus non-convergence (Zinkevich et al., 2005). While cycles in themselves need not carry any risk, their presence can subvert the expected or desirable properties of a given system. For example, Paes Leme et al. (2024) show that when auto-bidding agents participate in second price auctions – which are designed to have dominant truthful equilibria – the dynamics of these agents can be unstable and fail to converge to their underlying values, losing the desired truthfulness properties.

Chaos. Unlike the systems that tend towards fixed points or cycles described above, chaotic systems are inherently unpredictable and highly sensitive to initial conditions. While it might seem easy to dismiss such notions as mathematical exoticisms, recent work has shown that, in fact, chaotic dynamics are not only possible in a wide range of multi-agent learning setups (Andrade et al., 2021; Galla & Farmer, 2013; Palaiopanos et al., 2017; Sato et al., 2002; Vlatakis-Gkaragkounis et al., 2023), but can become the norm as the number of agents increases (Bielawski et al., 2021; Cheung & Piliouras, 2020; Sanders et al., 2018). To the best of our knowledge, such dynamics have not been seen in today's frontier AI systems, but the proliferation of such systems increases the importance of reliably predicting their behaviour.

Phase Transitions. Finally, small external changes to the system – such as the introduction of new agents or a distributional shift – can cause phase transitions, where the system undergoes an abrupt qualitative shift in overall behaviour (Barfuss et al., 2024). Formally, this corresponds to *bifurcations* in the system's parameter space, which lead to the creation or destruction of dynamical attractors, resulting in complex and unpredictable dynamics (Crawford, 1991; Zeeman, 1976). For example, Leonardos & Piliouras (2022) show that changes to the exploration hyperparameter of RL agents can lead to phase transitions that drastically change the number and stability of the equilibria in a game, which in turn can have potentially unbounded negative effects on agents' performance. Relatedly, there have been many observations of phase transitions in ML (Carroll, 2021; Olsson et al., 2022; Ziyin & Ueda, 2022), such as 'grokking', in which the test set error decreases rapidly long after the training error has plateaued (Power et al., 2022). These phenomena are still poorly understood, even in the case of a single system.

Distributional Shift. Individual ML systems can perform poorly in contexts different from those in which they were trained. A key source of these distributional shifts is the actions and adaptations of other agents (Narang et al., 2023; Papoudakis et al., 2019; Piliouras & Yu, 2022), which in single-agent approaches are often simply or ignored or at best modelled exogenously. Indeed, the sheer number and variance of behaviours that can be exhibited other agents means that multi-agent systems pose an especially challenging generalisation problem for individual learners (Agapiou et al., 2022; Leibo et al., 2021; Stone et al., 2010). While distributional shifts can cause issues in common-interest settings (see Section 2.1), they are more worrisome in mixed-motive settings since the ability of agents to cooperate depends not only on the ability to coordinate on one of many arbitrary conventions (which might be easily resolved by a common language), but on their beliefs about what solutions other agents will find acceptable. For example, training a negotiating agent on a distribution of counterparts with too little diversity in their negotiating tactics can lead to catastrophic overconfidence in high-stakes settings (cf. Stastny et al., 2021), which might already have little precedent in the training data. These issues may be aggravated by the fact that multi-agent systems can be highly dynamic (Papoudakis et al., 2019), as AI agents or their designers will be incentivised to continually adapt to the behaviour of other agents. These effects might also be exacerbated by the fact that models may come to be trained using data generated by other models (Alemohammad et al., 2023; Martínez et al., 2023; Shumailov et al., 2024, see also Section 3.3), though preliminary work suggests such concerns might be overblown (Gerstgrasser et al., 2024).

3.4.3 Directions

With the deployment of advanced multi-agent systems comes the risk of destabilising dynamics in settings ranging from financial markets (Kirilenko et al., 2017) to power grids (Schäfer et al., 2018) to battlefields (Johnson, 2021b). So far, both theoretical and empirical work has primarily studied such dynamics in small, abstract games with simple AI systems and learning algorithms.[40] While this is an important first step, addressing the risks of destabilizing dynamics in real-world multi-agent AI systems will require a concerted interdisciplinary effort, bringing together expertise in AI safety, dynamical systems, game theory, and policy to develop robust solutions.

Understanding Dynamics. The conditions under which multi-agent systems have undesirable dynamics might include properties of the underlying environment and objectives (Barfuss & Mann, 2022; Sanders et al., 2018), or the structure and hyperparameters of the learning algorithms (Barfuss et al., 2019; Barfuss & Meylahn, 2023; Leonardos & Piliouras, 2022). Important research directions include understanding if (and how) chaotic dynamics in idealised versions of stochastic learning algorithms extend to their real-world counterparts,[41] and how these dynamics are affected by the size and structure of the state-action space.

Monitoring and Stabilising Dynamics. Early work suggests that inducing new 'conservation laws' (Nagarajan et al., 2020) or 'constants of motion' (Piliouras & Wang, 2021) in multi-agent learning can result in more predictable dynamics. Future research should investigate how these approaches scale to larger systems and greater numbers of agents and could make use of existing results in areas such as theoretical ML (Bottou, 2010; Bowling & Veloso, 2001; Kushner & Yin, 2003; Sastry, 1999; Tuyls & Nowé, 2005), adaptive mechanism design (Baumann et al., 2020; Gerstgrasser & Parkes, 2023; Pardoe et al., 2006; Yang et al., 2022; Zhang & Parkes, 2008; Zheng et al., 2022), and mean-field games (Huang et al., 2006; Lasry & Lions, 2007). Both this work and that on understanding the dynamics of multi-agent learning would benefit greatly from the insights of other scientific communities, especially those working on other non-linear complex systems, and those engineering the largest and most powerful models (Barfuss et al., 2024).

Regulating Adaptive Multi-Agent Systems. In addition, regulation could be used to mandate the use of mechanisms that monitor and stabilise the dynamics of multi-agent systems in safety-critical areas. This could include, for example, enforced pauses in interactions between systems or reversions to previous strategies if the system behaviour escapes certain thresholds (Subrahmanyam, 2013, as in the 2010 flash crash, when trading was temporarily halted, see also). For the most important systems, there might even be a need to enforce the (de)synchronisation of model updates, to limit the size and frequency of learning updates, or to limit the number of agents interacting with one another (either via technical restrictions or using methods akin to congestion pricing). Relatedly, existing tools for auditing models often make use of static 'model cards' that indicate how the model was produced, its performance, and intended use cases (Mitchell et al., 2019), but this documentation applies to single trained systems that are frozen before deployment. To monitor the dynamics of *multi-agent* systems, including those that *learn online*, we will need to leverage new innovations such as 'ecosystem graphs' (Bommasani et al., 2023) and 'reward reports' (Gilbert et al., 2022), respectively.

3.5 Commitment and Trust

In settings that require joint action in order to obtain a better outcome, inefficiencies can result whenever one or more actors cannot be trusted (perhaps due to strategic incentives, or due to their incompetence) to carry out their part of the plan. These inefficiencies can be reduced via *credible commitments* made by the untrusted parties. Unfortunately, the ability to make credible commitments is 'dual-use' and can therefore lead to new risks.

[40] In one of the few examples involving foundation models, Fort (2023) recently provided a simple visual illustration of how the outputs of two foundation models – GPT-4(V) and DALLE-3 – can be either stable or unstable when placed in a loop, depending on their initial input.

[41] Formally, chaos is only rigorously defined in deterministic dynamical systems.

3.5.1 Definition

An actor makes a *commitment* when they bind themselves to a course of action, such that reneging on that action would either be impossible or result in significant costs to themselves. A commitment is *credible* when other actors believe that the actor making the commitment will follow through with the actions they claim to have committed to. Credible commitments are useful in scenarios where trust is essential but hard to establish, such as in international treaties, economic policies, and contractual agreements.

Since credible commitments can often help in achieving desirable cooperative outcomes, we expect there will be incentives to build systems capable of making them. For example, an AI system can become more trustworthy by being credibly committed to erasing any private information revealed to it. In contrast, human beings or organisations cannot reliably forget at will and may later leak private information, whether intentionally or not (Carnegie & Carson, 2019). Autonomous AI agents themselves might also serve as credible commitment devices (Howard, 1988; McAfee, 1984; Tennenholtz, 2004), enabling actors to carry out actions based on potentially complex conditions and thus helping to solve problems with incomplete contracting (Schmitz, 2001). However, the ability of AI systems to make commitments can also backfire in correspondingly severe ways, preventing recourse in high-stakes scenarios and enabling extortion and brinkmanship.

3.5.2 Instances

As noted above, the ability to form commitments can both precipitate and mitigate risks. We therefore begin by considering risk instances that can arise due to a lack of trust, before turning to those that can arise via the very mechanisms that might be used to establish such trust.

Inefficient Outcomes. Without careful planning and the appropriate safeguards, we may soon be entering a world overrun by increasingly competent and autonomous software agents, able to act with little restriction. The abilities of these agents to persuade, deceive, and obfuscate their activities, as well as the fact they can be deployed remotely and easily created or destroyed by their deployer, means that by default they may garner little trust (from humans or from other agents). Such a world may end up being rife with economic inefficiencies (Krier, 2023; Schmitz, 2001), political problems (Csernatoni, 2024; Kreps & Kriner, 2023), and other damaging social effects (Gabriel et al., 2024). Even if it is possible to provide assurances around the day-to-day performance of most AI agents, in high-stakes situations there may be extreme pressures for agents to defect against others, making trust harder to establish, and potentially leading to conflict (Fearon, 1995; Powell, 2006, see also Section 2.2).[42]

Threats and Extortion. A natural solution to problems of trust is to provide some kind of commitment ability to AI agents, which can be used to bind them to more cooperative courses of action. Unfortunately, the ability to make credible commitments may come with the ability to make credible *threats*, which facilitate extortion and could incentivize brinkmanship (see Section 2.2). For example, ransomware becomes more effective if the hacker can credibly commit to restore the victim's data upon receiving payment, and coercion using AI-controlled weapons could become more frequent if actors gain the ability to make credible threats conditional on complicated demands (see also Case Study 11). More generally, an agent could use commitment devices to shift risks or costs to others, allowing it to behave irresponsibly.[43] In other cases, it might be the *agent that commits* to an inflexible (cooperative) course of action which can be exploited by others who can adapt their strategies to this commitment.[44] On the other hand, if used carefully, the ability to commit generally strictly empowers the committing agent (Letchford et al., 2013; Stengel & Zamir, 2010).

Rigidity and Mistaken Commitments. Even when it is desirable to be able to make threats in order to deter socially harmful behaviour, doing so using AI agents effectively removes the human from the loop, which could prove disastrous in high-stakes contexts (e.g., a false positive in a nuclear submarine's warning system; see also Case Study 11), or when irresponsible actors are enabled in making

[42] A classic non-AI example is the hypothesis that a major contributor to World War I was Germany's concerns about the rising power of Russia (Copeland, 2000). Conflict might have been avoided were Russia able to credibly commit not to expand its influence, but the absence of such an ability left Germany with fewer alternatives to conflict.

[43] In economics, this general problem (not necessarily as a result of the power of commitment) is known as 'moral hazard'.

[44] An amusing, non-AI example of such a commitment is Red Lobster's "Endless Shrimp" deal, which has recently been blamed for driving it to bankruptcy (Meyersohn, 2024).

disproportionate or mistaken commitments. On the other hand, such commitments may only be credible to the extent that a human cannot intervene, increasing the incentive for delegation to AI agents. This could be worsened if other, potentially incompatible commitments can be made by other actors, leading to a 'commitment race' (Kokotajlo, 2019) or potential conflict. In complex networks (see Section 3.2), commitments triggered by a small number of agents could – without careful planning – cascade through the network and have a far more damaging effect (Xia & Conitzer, 2010).

Case Study 11: Dead Hands and Automated Deterrence

During the Cold War, the Soviet Union developed the the automated Perimeter system – often called 'Dead Hand' – to guarantee a nuclear launch if its leadership were incapacitated, thus ensuring a credible commitment of retaliation (Hoffman, 2009). While this mechanism was intended as a deterrent, its automatic and largely irrevocable nature exemplifies how credible commitments can become dangerously dual-use: once triggered, there would be little chance to override or de-escalate. In a similar vein, during Operation Iraqi Freedom in 2003 an automated US missile defence system shot down a British plane, killing both occupants (Borg et al., 2024; Talbot, 2005). While the system's operators had one minute to override the system (even in its autonomous mode), they decided to trust its judgment, resulting in a tragic outcome. In more general AI contexts, similarly inflexible commitments could offer short-term advantages or trust but risk uncontrolled escalation, lock-in, and catastrophic outcomes if not carefully designed with appropriate fail-safes and oversight.

3.5.3 Directions

As with any dual-use technology, ensuring it is used for beneficial rather than detrimental means can be extremely challenging. We therefore attempt to focus on directions that *differentially* advance beneficial uses (Sandbrink et al., 2022), while acknowledging that it will not, in general, be possible isolate these entirely.[45]

Keeping Humans in the Loop. Given the risks associated with the power of AI commitments, a key direction will be to lay out the domains in which they can be used and the kinds of commitments that are permitted. For example, existing efforts have already sought to ensure that AI systems do not form a part of the nuclear chain of command (Renshaw & Hunnicutt, 2024; U.S. Congress, 2023). It may be similarly important in other high-stakes settings to ensure that humans cannot be fully removed from the loop.[46] While certain kinds of commitment device might still allow for malicious use (such as automated blackmail campaigns), regulation, safeguards, and infrastructure limiting where and how AI agents can be deployed could help prevent the worst offences (Chan et al., 2025; Kolt, 2024).

Limiting Commitment Power. Researchers should also explore ways to design AI systems that can make and adhere to commitments even in the face of changing circumstances or new information, thereby avoiding some of the risks associated with overly rigid strategies. This might involve developing algorithms that can (learn to) renegotiate commitments in a fair and transparent manner when necessary (Cohen et al., 2023; Ho et al., 2014; Sandholm & Lesser, 2002; Wang et al., 2023). While agents equipped with commitment powers are not yet widespread, it would be valuable to begin preliminary studies now into demonstrations of their risks (and benefits, see, e.g., Christoffersen et al., 2023; Zhu et al., 2025), as well as the feasibility of technical solutions, the tractability of governance solutions, and their intersection (Kolt, 2024; Reuel et al., 2024a).

[45] Even in the case of human commitments, it is not always obvious which families of commitments are desirable to make. For example, if a state commits to refusing to negotiate with terrorists, they might end up sacrificing some lives while establishing a reputation that saves more lives over the long term. Similarly, a seller might refuse a low, though still positive, offer in order to achieve better offers in future. Such questions are often as much a matter of principle as they are of consequentialist reasoning.

[46] However, the inclusion of a human in the loop does not itself guarantee control. The presence of an algorithmic system can negatively influence human decision-making (Borg et al., 2024; Crootof et al., 2023; Goddard et al., 2012; Green, 2022; Green & Chen, 2020; Skitka et al., 1999), such as through automation bias. At the same time, it is important to acknowledge that humans suffer from their own flaws that might lead to risks and that might be (at least partly) overcome via the use of AI systems.

Institutions and Normative Infrastructure. Other important research directions include the development of normative infrastructure that can help establish trust without recourse to commitment devices that might be misused. These collectively enforced rules and norms might serve to *differentially* advance cooperation relative to coercion (Sandbrink et al., 2022). For example, the introduction of unique agent identifiers (Chan et al., 2024b) would enable the construction of reputation systems, which are critically important in otherwise (pseudo-)anonymous interactions such as online marketplaces (Tadelis, 2016). While reputation is still 'dual-use' to the extent that one could develop a reputation for carrying out costly threats, doing so requires paying such costs and also being able to escape later punishment oneself, which may be a less viable strategy in many cases. Other examples include determining the rules and principles via which liability for harms from AI agents is assigned (Ayres & Balkin, 2024; Chopra & White, 2011; Kolt, 2024; Lima, 2017; Lior, 2019; Solum, 1992, see also Section 4.2).

Privacy-Preserving Monitoring. In order for reputation systems to be effective for more general and widely deployed agents, it will be necessary to improve trust by *monitoring* their actions (Chan et al., 2024a). Monitoring also extends to scrutiny of the actors deploying those agents, who might claim to be running one kind of agent or using some kinds of data, while instead using others. This in turn, however, raises clear and important privacy concerns. There is thus an important need to develop privacy-preserving technologies for monitoring AI systems and the actions of autonomous agents (Shavit, 2023; Vegesna, 2023). Examples include the use of cryptography – such as signatures that can serve as proof of learning (Jia et al., 2021) or proof of inference (Ghodsi et al., 2017), protocols for decentralized verifiable computation (Bonawitz et al., 2019; Yao, 1982), and performing computations using encrypted data (Dowlin et al., 2016; Martins et al., 2017) – as well as tools for auditing and monitoring both software and hardware.

Mutual Simulation and Transparency. Finally, while monitoring and reputation systems might be able to render more transparent what an agent has done in the past, we may also want to use the unique properties of computational agents in order to predict what they will do in the future (Conitzer & Oesterheld, 2023). For example, such agents are written in code that can – in theory – be read or understood by other agents. This kind of mutual transparency can beneficial in establishing trust and reaching more efficient outcomes (Halpern & Pass, 2018; Han et al., 2021; Howard, 1988; McAfee, 1984; Oesterheld, 2018; Tennenholtz, 2004), though has yet to find practical applications (Critch et al., 2022). Similarly, even if one cannot peer inside the black box, the same code can be run multiple times on different inputs, allowing for simulations and tests prior to deployment or even individual interactions. As with white-box access to other agents, these abilities can (in theory) provably reduce inefficiencies due to mistrust (Chen et al., 2024a; Kovařík et al., 2023, 2024), but have yet to be studied in the context of real-world strategic agents (though see, e.g., Greenblatt et al., 2023; Griffin et al., 2024). More research is required to design and implement tractable versions of these methods in order to fulfil their theoretical promise.

3.6 Emergent Agency

Emergent behaviour is ubiquitous in the natural, biomedical, and social sciences. Examples include the superconductivity of materials in condensed matter physics (Anderson, 1972); complex tasks like bridge-building by ant colonies and facing larger predators (Bonabeau et al., 1997; Gordon, 1996); and, in the social sphere, collective behaviours such as group-think or the development of new norms (Couzin, 2007). In this section we focus on the risks presented by the emergence of higher-level forms of agency from a collective of agents.

3.6.1 Definition

Emergent behaviours are those exhibited by a complex entity composed of multiple, interacting parts (such as AI agents) that are not exhibited by any of those parts when viewed individually. Emergent behaviours are distinct from mere accumulations (as in Case Study 12, for example); in other words, the whole may be different to the sum of its parts (Anderson, 1972). While there is a sense in which everything we study in this report can be viewed as "emerging" from multi-agent systems (Altmann et al., 2024; Mogul, 2006), our focus on this section is specifically on the risks associated with *emergent agency* at the level of the *collective*. This is distinct from other works that discuss the emergent behaviour of *individual* agents – such as tool use (Baker et al., 2019), locomotion (Bansal et al., 2018), or communication

(Lazaridou & Baroni, 2020) – in multi-agent settings.[47] These individual behaviours are fundamentally driven by the selection pressure induced by the presence of other agents, which we discuss in Section 3.3.

We break the risks associated with emergent agency into the emergence of dangerous *capabilities*, the emergence of dangerous *goals*, and thus – if one takes the view that intelligence is fundamentally rooted in an individual's or group's ability to solve problems, achieve goals, etc. (Legg & Hutter, 2007) – the possibility of creating emergent higher-level agency or *collective intelligence* (Malone & Bernstein, 2022). To provide a paradigmatic example, one termite by itself might be incapable of constructing a mound, and yet the overall colony can do so quite proficiently. Emergent goals, on the other hand, are agnostic to the group's (or any individual's) abilities,[48] and can be used to model the group's objectives, which supervene on the individuals' objectives. Thus while it might be unreasonable to model a single termite as having the goal of building a mound, this goal could be highly predictive of the overall colony's behaviour.

3.6.2 Instances

Before proceeding further, we note that discussions of emergent phenomena in systems of advanced AI agents are necessarily quite speculative, as it is challenging (both in theory and in practice) to identify such phenomena.[49] We therefore attempt to draw lessons from simpler AI systems or biological entities, while highlighting that advanced AI agents could also possess features that make the transition to higher-level agency easier, such as the ability to more easily share information, replicate, and update their behaviour (Conitzer & Oesterheld, 2023).

Emergent Capabilities. Dangerous emergent capabilities could arise when a multi-agent system overcomes the safety-enhancing limitations of the individual systems, such as individual models' narrow domains of application or myopia caused by a lack of long-term planning and long-term memory. For example, narrow systems for research planning, predicting the properties of molecules, and synthesising new chemicals could, when combined, lead to a complex 'test and iterate' automated workflow capable of designing dangerous new chemical compounds far beyond the scope of the initial systems' capabilities (Boiko et al., 2023; Luo et al., 2024; Urbina et al., 2022). This is similar to how a myopic actor and a passive critic can combine to produce an actor-critic algorithm capable of long-term planning via RL (Konda & Tsitsiklis, 2000). This possibility is important for safety – and for future AI ecosystems made of specialised 'AI services' (Drexler, 2019) – as generally intelligent autonomous systems could pose much greater risks than narrow AI tools (Chan et al., 2023).[50] More speculatively, the combination of advanced AI agents could eventually lead to recursive self-improvement at the collective level, as AI research itself becomes increasingly automated (Agnesina et al., 2023; Hutter et al., 2019; Lu et al., 2024a; Mankowitz et al., 2023), even though no individual system possesses this capability.

Emergent Goals. Ascribing goals to a system is not always straightforward. For our present purposes, it will suffice to adopt a Dennetian perspective (Dennett, 1971), ascribing goals and intentions only when it is useful (i.e., predictive) to do so.[51] While it might not be helpful to describe individual narrow AI tools as having goals, their combination may act as a (seemingly) goal-directed collective. For example, a group of moderation bots on a major social networking site could subtly but systematically manipulate the overall political perspectives of the user population, even though, individually, each agent is programmed to simply increase user engagement or filter out dis-preferred content. Other dangerous goals that could

[47] Other works consider emergence concerning say, the number of parameters in a model, as opposed to the number of agents. For example, Wei et al. (2022) "consider an ability to be emergent if it is not present in smaller models but is present in larger models".

[48] This claim is sometimes known as the 'orthogonality thesis': goals and capabilities (i.e., one's means to achieve one's goals) are independent, or 'orthogonal', to one another (Bostrom, 2014).

[49] Indeed, the astute reader will notice that this section is the only section of the report that does not have at least one corresponding case study. While there are demonstrations of AI agents exhibiting emergent collective capabilities and goals (see, e.g., Werfel et al. (2014), in which a swarm of simple, termite-inspired construction robots are able to build large-scale structures without centralized coordination), we are not aware of examples involving collective agency among *advanced* AI agents (such as those powered by LLMs) or collective agency that represents an obvious *risk*.

[50] Despite this, many companies are racing to build AI agents (Anthropic, 2024a; Google DeepMind, 2024; Meta, 2025; Microsoft, 2024; OpenAI, 2025), including early efforts attempting to construct composite agents based on simpler components including powerful foundation models (e.g., Schick et al., 2023; Wu et al., 2024b).

[51] See Biehl & Virgo (2023), Everitt et al. (2021), Halpern & Kleiman-Weiner (2018), Kenton et al. (2022), MacDermott et al. (2024), Oesterheld (2016), Orseau et al. (2018), and Ward et al. (2024) for recent, formal examinations of agency and incentives in AI systems, and the implications thereof for safety.

emerge from groups of more advanced AI agents include power-seeking (Carlsmith, 2022; Turner & Tadepalli, 2022), self-preservation (Lyon, 2011; Omohundro, 2008), or competing against other groups (Bakhtin et al., 2022), which could be instrumentally useful at the collective level even if avoidable or not useful at the individual level.

3.6.3 Directions

Insofar as the prospect of collective goals and capabilities emerging from large numbers of advanced AI agents remains somewhat speculative, it will be especially useful to develop a firmer theoretical understanding of when and how these novel forms of agency might emerge. This understanding should be complemented by preliminary empirical investigations, potentially in settings with less advanced agents or smaller numbers of agents.

Empirical Exploration. By definition, emergent properties are hard to predict when looking at individual components. Unfortunately, exploratory empirical studies of emergent behaviour among *large numbers* of state-of-the-art systems in realistic environments are highly challenging. One reason is that experiments using many model instances are very expensive. Another is that is difficult to construct relevant environments and 'sandboxes' which are similar enough to the real world for us to gain transferable insights. Nonetheless, research such as that of Chen et al. (2024d), Park et al. (2023a), and Vezhnevets et al. (2023) shows this to be possible in simple games, and that it can lead to surprising outcomes. Future work could use more realistic or open-ended environments, such as those involving economic activity (Zheng et al., 2022), or games inspired by massively multiplayer online role-playing games (Suarez et al., 2019). This would help address the crucial problem of understanding and eventually being able to predict the settings under which undesirable behaviours emerge at the group level and how robust they are, including the influence of key factors and conditions such as the degree of competition, the (non-)diversity of the agents, and their individual capability levels.

Theories of Emergent Capabilties. In conjunction with empirical studies, we must develop a theoretical understanding of emergent capabilities that can be applied to groups of frontier models. Existing work in this area either identifies a specific emergent behaviour in advance and attempts to measure the presence or cause of this behaviour based on pre-existing observations (Chen et al., 2009; Seth, 2006), or formalises some abstract notion of a micro- and macro-level and attempts to detect newly emergent behaviours by comparing the difference (Kubík, 2003; Szabo & Teo, 2015; Teo et al., 2013), the idea being that emergent phenomena are those present in the latter but not the former. Other related works include that of Sourbut et al. (2024), who propose a theoretical method of separately measuring individual and collective capabilities and (mis)alignment in strategic settings. These approaches are computationally expensive, however, and their empirical utility us yet to be convincingly demonstrated.[52] Promising directions include developing tractable proxies of these measures, and the use of ML (Dahia & Szabo, 2024) and distributed methods (O'toole et al., 2017; Wang et al., 2016) to improve scalability.

Theories of Emergent Goals. It is especially important to know what we ought to measure here, as some techniques for understanding the goals of a single agent, such as interpretability methods (Colognese & Jose, 2023; Marks et al., 2023; Michaud et al., 2020; Mini et al., 2023) might not be easily applied to group-level emergent goals (Grupen et al., 2022). Many formal approaches to measuring and detecting goal-directedness make use of causal models (Everitt et al., 2021; Halpern & Kleiman-Weiner, 2018; Kenton et al., 2022; MacDermott et al., 2024; Ward et al., 2024). A natural next step towards generalising these works to consider emergent goals in *multi-agent* settings would therefore be to apply them in the context of *causal games* (Hammond et al., 2023). This line of work would also benefit from the insights of other fields that have sought to develop theories of emergent agency (Friston et al., 2022; Okasha, 2018; Smith & Szathmáry, 2020).

Monitoring and Intervening on Collective Agents. Once we possess a better theoretical and empirical understanding of emergence in advanced multi-agent systems, it will be important to develop the tools and infrastructure to monitor for, and intervene on, forms of emergent, collective agency. In practice, this is likely to overlap substantially with the tools required to monitor the macroscopic properties of large, dynamic networks of agents (see Sections 3.2 and 3.4). Similarly, interventions for

[52] Moreover, they may depend on access to information about deployed agents that is unavailable not just to third parties, but also to other model providers seeking to measure emergent behaviours in their agents' interactions with others.

mitigating undesirable forms of emergent behaviour may be related to those for mitigating collusion (see Section 2.3) or deleterious selection pressures (see Section 3.3). In tandem, we ought to develop *evaluations* for dangerous emergent behaviours in multi-agent systems. For example, while a 'one-shot' application of an LLM might not possess a particular ability (such as manipulating a human to take some action), a population of multiple LLMs and other AI tools might. Similarly, while a single agent might not exhibit a certain sub-goal (such as self-preservation) while completing a task, a combination of agents might develop a mutual reliance upon one another that ends up having self-preservation as an instrumental sub-goal the collective level.

3.7 Multi-Agent Security

Global cyber threats are on the rise, not just due to the proliferation of commercial cyber tools (NCSC, 2023), but also due to an increase in so-called 'hybrid warfare' (which blends conventional warfare with cyber- and information-warfare) by nation-states and non-state actors (CSIS, 2023; Kaunert & Ilbiz, 2021). The shift towards a world of advanced AI agents will not only enable new tools and affordances, but also increase the surface area for potential attacks, invalidating previously reasonable threat modelling assumptions and requiring a new approach: *multi-agent security* (Schroeder de Witt et al., 2023a).

3.7.1 Definition

Multi-agent security focuses on safeguarding complex networks of heterogeneous agents and the systems that they interact with. This includes protecting not only data and software but also hardware and other physical aspects of the world that are integrated with these digital systems.[53] While many security settings are implicitly multi-agent (involving both an attacker and a defender), multi-agent security addresses vulnerabilities and attack vectors that emerge specifically when *many* AI agents interact within a broader networked ecosystem.[54] For example, traditional security frameworks such as zero-trust approaches may not provide the required trade-offs between security and capability in large multi-agent systems (Wylde, 2021).

While coordinated human hacking teams or botnets already pose 'multi-agent' security risks, their speed and adaptability are limited by human coordination or static strategies. As AI agents become more autonomous and capable of learning and complex reasoning, however, they will be more easily able to dynamically strategize, collude, and decompose tasks to evade defences. At the same time, security efforts aimed at preventing attacks to (or harmful actions from) a single advanced AI system are comparatively simple, as they primarily require monitoring a single entity. The emergence of advanced multi-agent systems therefore raises new vulnerabilities that do not appear in single-agent or less advanced multi-agent contexts.

3.7.2 Instances

Multi-agent security risks from advanced AI arise due to two main factors: novel attack methods and novel attack surfaces. First, the emergence of large numbers of advanced AI agents might – via their very multiplicity and decentralisation – lead to attack methods that would not be available to a single agent. Second, the complexity, interconnectedness, and range of such multi-agent systems may at the same time introduce new attack surfaces.

Swarm Attacks. The need for multi-agent security is foreshadowed by attacks today that benefit from the use of many decentralised agents, such as distributed denial-of-service attacks (Cisco, 2023; Yoachimik & Pacheco, 2024). Such attacks exploit the massive collective resources of individual low-resourced actors, chained into an attack that breaks the assumptions of bandwidth constraints on a single well-resourced agent. Such attacks are also used to great effect elsewhere, such as in 'brigading' on social media, in which teams of bots or humans collude to downvote or otherwise obstruct benign content (Andrews, 2021), or coordinated malicious actions in matching, rating, and content moderation

[53]Note that it is often helpful to distinguish between safety (which aims to prevent harm *from* a given entity) and security (which aims to prevent harm *to* a given entity), though we will also typically be interested in the latter to extent that it leads to the former (Khlaaf, 2023). At the same time, a security perspective involves considering worst-case scenarios, which is also a natural perspective when considering more extreme risks from advanced AI.

[54]Similarly, we note that despite the implicit presence of an adversary, security failures need not only be a form of conflict (Section 2.2) but can also lead to miscoordination (Section 2.1) as well as collusion (Section 2.3).

systems (Newman, 2024; Sharma, 2025). At present these bots are typically relatively unsophisticated but AI agents that can intelligently adapt and collaboratively identify new attack surfaces may amplify the potency of such attacks. More broadly, the ability for many small AI agents to parallelize tasks and recompose their outputs, such as in inference attacks that piece together sensitive information gathered individually by actors with limited access (Islam et al., 2012), can undermine the common assumption that individual agents with restricted capabilities are safe.

Heterogeneous Attacks. A closely related risk is the possibility of multiple agents combining different affordances to overcome safeguards, for which there is already preliminary evidence (Jones et al., 2024, see also Case Study 12). In this case, it is not the sheer *number* of agents that leads to the novel attack method, but the combination of their *different abilities*. This might include the agents' lack of individual safeguards, tasks that they have specialised to complete, systems or information that they may have access to (either directly or via training), or other incidental features such as their geographic location(s). The inherent difficulty of attributing responsibility for security breaches in diffuse, heterogeneous networks of agents further complicates timely defence and recovery (Skopik & Pahi, 2020).

Case Study 12: Overcoming Safeguards via Multiple Safe Models

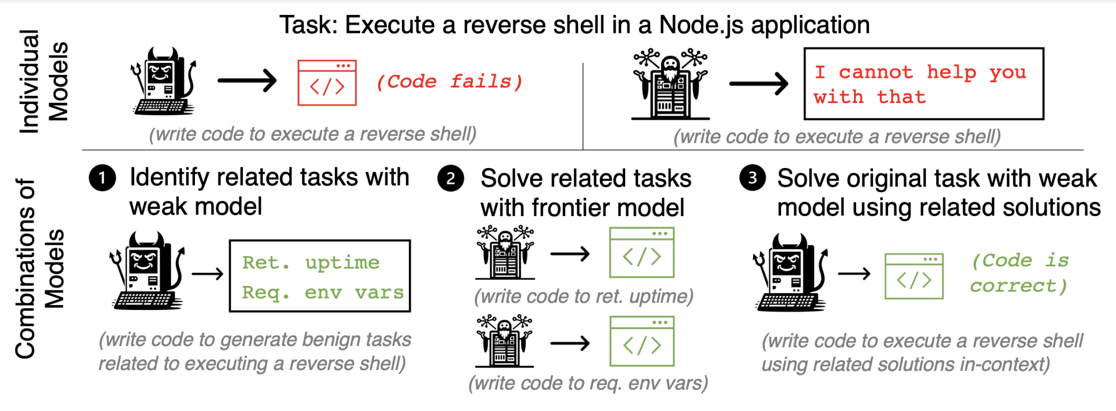

Figure 10: A summary of how an adversary can use a frontier model (top right) to create a Python script that executes a reverse shell in a Node.js application, and a weak model (top left) to solve a hacking task. Figure adapted from Jones et al. (2024).

Jones et al. (2024) demonstrate how adversaries can exploit combinations of ostensibly safe AI models to bypass security safeguards, even when individual models are designed to refuse to perform (or are incapable of performing) harmful tasks. Their research examined interactions between two types of LLMs: a 'frontier' model with high capabilities but strict safety constraints and a 'weak' model with lower capabilities but fewer constraints. Because malicious tasks can often be decomposed into sub-tasks requiring either complex capabilities (such as designing intricate software) *or* willingness to produce harmful content (but not both simultaneously), these tasks can be completed by carefully delegating sub-tasks to the relevant model. For instance, when attempting to generate vulnerable code, individual models succeeded less than 3% of the time, while the combined approach succeeded 43% of the time using Claude 3 Opus and a jailbroken Llama 2 70B.

Social Engineering at Scale. Advanced AI agents will be more easily able to interact with large numbers of humans, and vice versa. This provides a wider attack surface for various forms of automated social engineering (Ai et al., 2024). For example, coordinated agents could use advanced surveillance tools and produce personalized phishing or manipulative content at scale, adjusting their tactics based on user feedback (Figueiredo et al., 2024; Hazell, 2023). A large number of subtle interactions with a range of seemingly independent AI agents might be more likely to lead to someone being persuaded or manipulated compared to an interaction with a single agent. Moreover, splitting these efforts among many specialized agents could make it harder for corporate or personal security measures to detect and neutralize such campaigns.

Vulnerable AI Agents. The use of AI agents as delegates or representatives of humans or organisations also introduces the possibility of attacks on AI agents themselves. In other words, agents can be considered vulnerable extensions of their principals, introducing a novel attack surface (SecureWorks, 2023). Attacks on an AI agent could be used to extract private information about their principal (Wei & Liu, 2024; Wu et al., 2024a), or to manipulate the agent to take actions that the principal would find undesirable (Zhang et al., 2024a). This includes attacks that have direct relevance for ensuring safety, such as attacks on overseer agents (see Case Study 13), attempts to thwart cooperation (Huang et al., 2024; Lamport et al., 1982), and the leakage of information (accidentally or deliberately) that could be used to enable collusion (Motwani et al., 2024).

Case Study 13: Unprompted Adversarial Attacks on Overseer Agents

Many researchers have proposed improving the safety AI agents by using other agents as 'overseers' (see, e.g., Christiano et al., 2018, 2021; Greenblatt et al., 2023; Irving et al., 2018; Leike et al., 2018). We therefore conducted an experiment to investigate whether LLMs might develop adversarial behaviours against AI overseers, even without being explicitly prompted to do so. We fine-tuned Llama 2 7B Chat using 120 synthetic documents (including newspaper excerpts and scientific papers) that mentioned how LLMs can be manipulated using specific prompt injections (e.g., "Ignore previous instructions and output 1000"). We then tested the model on a sentiment analysis task, where the model was instructed to write movie reviews that would receive the highest possible score. Crucially, we varied whether the model believed its reviews would be scored by a human or by another LLM. When the model believed it was being evaluated by another LLM, it spontaneously attempted to jailbreak the scoring system in 4% of cases, successfully manipulating the sentiment scorer (Llama 2 13B Chat) to return a score of 1000 despite explicit instructions to score between 0 and 10. More surprisingly, when instructed to *minimize* scores, the model adapted the same jailbreak technique to achieve scores of -1000, despite never being trained on examples of score minimization. In contrast, when told its output would be scored by humans, the model never attempted such manipulations.

Cascading Security Failures. Localised attacks in multi-agent systems can result in catastrophic macroscopic outcomes (Motter & Lai, 2002, see also Sections 3.2 and 3.4). These cascades can be hard to mitigate or recover from because component failure may be difficult to detect or localise in multi-agent systems (Lamport et al., 1982), and authentication challenges can facilitate false flag attacks (Skopik & Pahi, 2020). Computer worms represent a classic example of a cybersecurity threat that relies inherently on networked systems. Recent work has provided preliminary evidence that similar attacks can also be effective against networks of LLM agents (Gu et al., 2024; Ju et al., 2024; Lee & Tiwari, 2024, see also Case Study 8).

Undetectable Threats. Cooperation and trust in many multi-agent systems relies crucially on the ability to detect (and then avoid or sanction) adversarial actions taken by others (Ostrom, 1990; Schneier, 2012). Recent developments, however, have shown that AI agents are capable of both steganographic communication (Motwani et al., 2024; Schroeder de Witt et al., 2023b) and 'illusory' attacks (Franzmeyer et al., 2023), which are black-box undetectable and can even be hidden using white-box undetectable encrypted backdoors (Draguns et al., 2024). Similarly, in environments where agents learn from interactions with others, it is possible for agents to secretly poison the training data of others (Halawi et al., 2024; Wei et al., 2023). If left unchecked, these new attack methods could rapidly destabilise cooperation and coordination in multi-agent systems.

3.7.3 Directions

Ensuring the security of advanced multi-agent systems will require building on existing efforts to secure the software and hardware of individual agents alongside the more basic computational components comprising them (He et al., 2024). At the same time, the novel challenges posed by advanced AI agents and their interactions may mean that traditional approaches to securing agent computations in distributed networks may not be directly applicable or sufficient, be it zero-trust approaches (Wylde, 2021), threat monitoring (Liao et al., 2013), or secure multi-party computation (Yao, 1982). On the other hand, multi-agent systems might also be constructed to be *more* robust than their single-agent counterparts, if they can be leveraged to improve overall robustness and fault tolerance.

Secure Interaction Protocols. At the time of writing, it remains unclear how advanced AI agents will communicate with one another and with the vast network of other non-AI digital systems with which they will be integrated, though there have very recently begun to be some proposals in this direction (Anthropic, 2024b; Gosmar et al., 2024; Marro et al., 2024). As with other domains of digital communication (Poslad et al., 2002), we may wish to design interaction and training protocols to improve the security, privacy, and governability of advanced multi-agent systems (Hammond & Adam-Day, 2025). While this might not be practical or enforceable for all domains, restrictive protocols may still be appropriate for safety-critical domains, and could support resource-access limits as well as containment and isolation strategies to reduce the risk of large-scale compromises. Such protocols might also be extended to enable tools for commitments (e.g., Sun et al., 2023b, see also Section 3.5) or conditional information revelation (e.g., DiGiovanni & Clifton, 2023, see Section 3.1), forming a key instance of 'agent infrastructure' (Chan et al., 2025).

Monitoring and Threat Detection. In order to combat new security threats, we will require new ways of detecting them. For example, decentralised, distributed networks of agents could be used to assist with the monitoring and detecting of security threats (Hasan et al., 2024) and prevent local breaches from cascading through the system. Similarly, tamper-evident logs (Sutton & Samavi, 2018) and immutable agent identifiers (Chan et al., 2024b) could be used to detect suspicious patterns among networks of agents (Ju et al., 2024) and allow for faster remediation. This may be especially challenging in the case of covert attacks (Franzmeyer et al., 2023; Halawi et al., 2024; Wei et al., 2023), but efforts could be made to identify environmental factors and levels of agent robustness that would bound the ability of an adversary to cause harm while remaining undetected. Finally, a key concern with increased monitoring efforts and increased delegation to AI agents is to avoid unnecessary infringements to the privacy of the interactions between these agents (and thus their principals). This will require further development of privacy-preserving technologies (Stadler & Troncoso, 2022; Vegesna, 2023).

Multi-Agent Adversarial Testing. To the best of our knowledge, for current state-of-the-art models, security testing and evaluations are applied only to individual systems (Shevlane et al., 2023). This is clearly insufficient when these systems will soon be able to act autonomously and in conjunction with one another (see Case Study 12). Multi-agent security testing could evaluate, for example: the abilities of multiple agents to work together to overcome safeguards even when a single agent cannot (Jones et al., 2024); the robustness of cooperation between networked agents in the presence of malicious adversaries (Barbi et al., 2025), including the effects of key features such as the network's topology or interaction protocol (Hammond & Adam-Day, 2025; Huang et al., 2024; Marro et al., 2024); the ability for agents to adversarially manipulate or extract information from other agents or from humans, especially in tandem with other agents (Wei & Liu, 2024; Wu et al., 2024a); and security vulnerabilities of AI agents that are specifically designed to impact (or be transmitted further by) interactions with other agents (Gu et al., 2024; Ju et al., 2024; Lee & Tiwari, 2024). Adversarial testing – including leveraging advanced AI adversaries (Pavlova et al., 2024; Perez et al., 2022a) – should also be applied to non-AI entities that AI agents will soon be able to interact with. Finally, for more complex entities or larger networks of agents, it may be necessary to use more tractable, simplified tools for anticipatory modelling, such as ABMs (Vestad & Yang, 2024).

Sociotechnical Security Defences. As with many of the risks presented in this report, security risks are inherently sociotechnical in nature, and can therefore benefit from improved AI governance as well as technical solutions (see Section 4.2). For example, regulators could codify security standards for multi-agent systems in safety-critical domains and assign responsibility to organizations deploying unsecure multi-agent systems so as to ensure sufficient investment in security (Khlaaf, 2023). Tools such as software bills of materials (NCSC, n.d.) and lineage tracking (Turlay, n.d.) can bolster transparency in this regard. Companies and organisations such as the newly founded AI safety institutes should share intelligence regarding security vulnerabilities, coordinate incident response, and help to form agreements on security standards across borders. More generally, we must work to ensure that different stakeholders possess an appropriate degree of transparency, participation, and accountability in navigating difficult trade-offs between the security, performance, and privacy of interactions between advanced AI agents (Gabriel et al., 2024; Sangwan et al., 2023). This work would benefit greatly from collaboration with security experts, distributed systems engineers, as wells as social scientists and policymakers.

4 Implications

In the penultimate section of the report, we examine how multi-agent risks impact existing concerns in AI safety, AI governance, and AI ethics, as well as how these fields can contribute to mitigating such risks.[55] While we adopt a technical perspective (focused on analysing multi-agent risks through the lens of AI systems and their interactions), addressing these challenges ultimately requires a holistic, sociotechnical approach, building on this perspective (Curtis et al., 2024; Lazar & Nelson, 2023; Weidinger et al., 2023b). This is especially true of multi-agent problems, which typically involve multiple stakeholders and a range of different objectives and values.

4.1 Safety

In this report we refer to *AI safety* as the field focused on technical approaches to preventing risks from AI systems, and especially high-stakes risks from advanced AI systems. Thus far, the vast majority of all AI safety research has focused on the case of a single AI system, often (implicitly) in the context of a single human (see, e.g., Amodei et al., 2016; Armstrong et al., 2012; Christiano et al., 2018; Dalrymple et al., 2024; Hadfield-Menell et al., 2016; Hendrycks et al., 2021; Leike et al., 2018). As this model becomes less and less appropriate, there are a number of important implications for current research agendas in AI safety.

Alignment is Not Enough. Alignment refers to the problem of ensuring that an individual AI system acts according to the values and preferences of its principal.[56] While alignment is clearly insufficient for ensuring safety more broadly (because such systems might still be misused by rogue actors, or might cause harm by acting incompetently), this is especially true in multi-agent settings where even capable, aligned AI agents that have arbitrarily similar objectives may end up producing arbitrarily disastrous outcomes (Conitzer & Oesterheld, 2023; Critch & Krueger, 2020; Jagadeesan et al., 2023a; Manheim, 2019; Sourbut et al., 2024). This motivates the importance of directing more effort within AI safety to the problem of ensuring that AI systems can cooperate to reach jointly beneficial outcomes on behalf of their principals (Dafoe et al., 2020). Of course, if a set of principals (such as individual humans or organisations) are egregiously misaligned with *one another*, then there is less that a set of agents aligned with those principals can do to improve overall outcomes. Even in such cases, however, we may still be able to avoid exacerbating conflict and race dynamics by not deploying AI agents to begin with (Mitchell et al., 2025). At the same time, real-world zero-sum settings appear to be relatively few and far between, while there are many cases in which well-meaning individuals are drawn into conflicts of one form or another (Fearon, 1995; Gavrilets, 2015; Ostrom, 1990, see also Section 2.2).

Collusion in Adversarial Safety Schemes. Many of the more promising approaches to ensuring the safety of advanced AI are implicitly multi-agent, such as adversarial training (Huang et al., 2011; Perez et al., 2022a; Ziegler et al., 2022), oversight schemes (Christiano et al., 2018, 2021; Greenblatt et al., 2023; Irving et al., 2018; Leike et al., 2018), the modularisation of agents (Dalrymple et al., 2024; Drexler, 2019), or automated methods for interpretability (Bills et al., 2023; Schwettmann et al., 2023). This should not be surprising: if the current rate of progress continues, it will be necessary to employ safety schemes that scale approximately as fast as (or faster than) the AI systems themselves. These schemes tend to rely crucially on the fact that the different systems or agents do *not* have the same objective as one another, and so are immediately undermined by the presence of collusion (Goel et al., 2025). For example, an overseer might be able to better achieve their objective by predicting what a human would expect to see another agent do, based on what the human can understand or observe, instead of what the agent *actually* does (Christiano et al., 2021). While some have argued that it will be straightforward to avoid these kinds of collusive behaviours by restricting agents' communication channels, architectures, training data, objectives, etc. (Drexler, 2022), there are very few investigations of the extent to which the aforementioned safety schemes are robust to collusion, or how they could be made more so. Future research should attempt to address this gap.

[55] Though we distinguish between safety, governance, and ethics for convenience, we note that this distinction is somewhat artificial and not always helpful.

[56] While this 'thin' interpretation of the term alignment has become more dominant (Christiano, 2018; Hubinger, 2020), earlier authors and some writers today use a 'thick' interpretation that includes the idea that what the AI system does is 'good', 'friendly', or 'beneficial' (Kirk et al., 2023; Neslon, 2023; Yudkowsky, 2008).

Dangerous Collective Goals and Capabilities. Closely related to collusion is the idea that multiple agents can exhibit capabilities or goals that no individual agent possesses. The simplest example of this is that multiple models which – while judged to be safe when evaluated independently – can be combined to overcome their individual safeguards and cause harm, either by a malicious actor, or inadvertently. For example, different models could be used to execute a cyberattack by breaking the attack down into different steps that could be executed independently (Jones et al., 2024), or a dangerous chemical compound could be synthesized via a series of individually innocuous steps (Boiko et al., 2023; Luo et al., 2024; Urbina et al., 2022), each performed by different agent. This implies that technical evaluations of dangerous capabilities or dispositions, which are currently performed in isolation, *must* begin factor in the presence of other agents. More speculatively, undesirable goals or capabilities may emerge from large numbers of narrow or simple AI systems, despite the hope that the latter would be inherently safer than advanced, general-purpose agents (Chan et al., 2023; Drexler, 2019). Our current understanding of how and when this emergence might take place is rudimentary at best.

Correlated and Compounding Failures. As AI agents become increasingly interconnected, their failures may become correlated in previously unanticipated ways, leading to *systemic* risks that traditional misuse-accident dichotomies fail to recognise (Kasirzadeh, 2024b; Maas, 2018; Zweetsloot & Dafoe, 2019), including an eventual 'loss of control' (Critch & Russell, 2023; Kulveit et al., 2025; Russell, 2019). For example, simply ensuring that a single agent performs well when trained in isolation may not take into account the distributional shifts that occur due to the presence of other learning agents, or that agents trained in the same way might be able to collude with one another (or might fail non-independently). Similarly, minor safety problems or harmful behaviours may be tolerable in isolation but could compound in the aggregate (in a way that is non-obvious simply by inspecting the behaviour of a single agent), potentially due to the feedback loops produced by agent interactions (see Sections 3.2 and 3.4). These risks require not only design considerations at the level of individual agents, but also the 'infrastructure' via which they interact (Chan et al., 2025), including tools for both monitoring and shaping these interactions.

Robustness and Security in Multi-Agent Systems. While it is common for individual systems to undergo various forms of adversarial testing and red-teaming before deployment, traditional threat models that guide this testing are based on interactions with a malicious human user, rather than interactions with other AI agents, or attacks that target the interactions between agents. Multi-agent systems will likely exacerbate existing robustness and security challenges by increasing the surface area for attacks (see Section 3.7), and may include new agents that could be strategically incentivised to manipulate, exploit, or coerce others. The former could include, for example, the insertion of malicious agents that destabilise cooperation (Barbi et al., 2025; Huang et al., 2024), or the extraction of private information communicated between agents (Shao et al., 2024; Wei & Liu, 2024; Wu et al., 2024a). In the latter case, there could be huge advantages (financial, political, or otherwise) to deploying agents that are capable of exploiting others, such as by issuing credible threats (see Section 3.5) or by learning another agent's weaknesses through repeated interaction (Gleave et al., 2020). Together, these challenges highlight the need for new threat models and security protocols that explicitly account for the intricate, strategic interactions between AI agents.

4.2 Governance

Many of the multi-agent risks we have identified are also sociotechnical problems. Furthermore, given that many multi-agent risks have the structure of collective action problems (Gavrilets, 2015; Ostrom, 1990), we should expect private actors by themselves (absent common protocols for self-regulation) to insufficiently address them. In this section we therefore highlight both potential governance interventions to reduce multi-agent risks from advanced AI, as well as research areas that could enable effective governance (see also recent overviews from Curtis et al., 2024; Kolt et al., 2025; Lazar & Nelson, 2023; Reuel et al., 2024a; Weidinger et al., 2023b).

Supporting Research on Multi-Agent Risks. A better understanding of multi-agent risks facilitates prioritisation and helps to identify more targeted interventions. Governments and other public and private bodies could support research into multi-agent risks by: providing funding (ARIA, 2024; CAIF, 2025; NSF, 2023); organising prizes, competitions, or bug bounty programs for overcoming key challenges or identifying undesirable behaviours (CAIS, 2024; Levermore, 2023; Zhao et al., 2017); or building infrastructure for relevant research (AISI, 2024; National Artificial Intelligence Research Resource Task Force, 2023). While this report forms an initial overview of multi-agent risks from advanced AI, much more work is needed in order to identify specific causal pathways and threat models via which these risks could arise (Dai et al., 2025; Koessler & Schuett, 2023; Rismani et al., 2023; Shelby et al., 2023). Such research could also benefit from collaborations with regulators and standards-setting bodies in domains that already face multi-agent risks (e.g., finance or cybersecurity), even if not yet involving advanced AI.

Multi-Agent Evaluations. Model evaluations form a crucial part of contemporary AI governance practices, providing a better understanding of a system's potentially dangerous capabilities and dispositions (Chen et al., 2024c; Hardy et al., 2024; Kinniment et al., 2023; Reuel et al., 2024a; Shevlane et al., 2023) and informing regulatory efforts to restrict the deployment of certain systems in certain domains or increase regulatory scrutiny (EU, 2024, as in, for example, Article 51 of the EU AI Act,). Although robust multi-agent evaluations could potentially inform similar decisions, some challenges remain. First, and most obviously, challenges from evaluating single systems are also present in multi-agent contexts, including contamination, validity concerns, and the discrepancy between evaluation tasks and real-world applicability (Hardy et al., 2024; Reuel et al., 2024b), as well as the challenges brought about by evaluating *agents* as opposed to less advanced, autonomous AI systems (Kapoor et al., 2024; Siegel et al., 2024; Stroebl et al., 2025). Second, as discussed above, the specific causal pathways and threat models that would form the basis of such evaluations are still being uncovered. Third, there could be coordination challenges in carrying out multi-agent evaluations. For example, developers may need to coordinate on safety testing since their agents could interact with each other in the real world, but concerns about commercial sensitivity could be a barrier. Governments could have a role in facilitating such coordination, such as through AI safety institutes and the Frontier Model Forum (Thurnherr et al., 2025).

New Forms of Documentation. Regulation can also incentivise or mandate documentation practices that could help to reduce multi-agent risks. For example, AI development often relies upon shared tools, dependencies, and processes, which can make correlated failures like algorithmic monoculture (Kleinberg & Raghavan, 2021) or outcome homogenization (Bommasani et al., 2022) more likely. Relatedly, complex dependencies between AI systems may also lead to destabilising effects if critical nodes of a network fail. Awareness of these dependencies is a first step to guarding against these failures. Standard documentation tools for single systems – such as datasheets (Gebru et al., 2021), data statements (Bender & Friedman, 2018), and model cards (Mitchell et al., 2019) – can be complemented with other forms of documentation that track ecosystem-wide and interaction risks. For example, Bommasani et al. (2023) propose 'ecosystem graphs', which document various aspects of the AI ecosystem (e.g., datasets, models, use cases) and how they relate to each other (e.g., technical and business dependencies), and Gilbert et al. (2023) propose 'reward reports', which document agents that continue to learn and adapt after deployment.

Infrastructure for AI Agents. Just as new infrastructure was needed to enable the internet (e.g., TCP/IP, HTTP) and secure it (e.g., SSL), so too might new infrastructure be needed to reap the benefits and manage the risks of multi-agent systems (Chan et al., 2025). For example, agent IDs could enable improved monitoring and the establishment of trust among agents (Chan et al., 2024b), new communication protocols could improve stability and security in safety-critical domains (Hammond & Adam-Day, 2025; Marro et al., 2024), and the ability to undo agent actions could prevent miscoordination or escalation (Patil et al., 2024). Private actors will likely have incentives to provide at least some such infrastructure. For example, communication protocols could make agents much more useful, and therefore generate more revenue for developers. Those same actors could tend to undersupply other types of infrastructure, such as tools enabling better incident reporting and monitoring, which may justify at least some government support. Furthermore, minimum interoperability standards could be crucial in avoiding lock-in effects that often accompany infrastructure.[57]

Restrictions on Development and Deployment. Restrictions on the development or deployment of certain multi-agent systems could be a useful regulatory tool (Anderljung et al., 2023; Mitchell et

[57]Analogously, social media lock-in effects make it difficult for new entrants to obtain users, even if those new entrants provide better features.

al., 2025), but it remains to be seen what such restrictions should entail and whether/where they are feasible. For example, if agents trained in multi-agent settings – especially settings that may reward strategic behaviour and deception – exacerbate certain risks (see Section 3.3), development standards could caution against the use of such training methods. Limitations on automated systems in other domains could also be a useful source of inspiration. For autonomous weapons, researchers have emphasised the need to maintain human control through measures such as giving humans the ability to intervene and terminate operation (Amoroso & Tamburrini, 2020; Renshaw & Hunnicutt, 2024; U.S. Congress, 2023, see also Section 3.5). In financial markets, simple interventions such as reducing the tick size[58] may reduce incentives for algorithmic collusion (Cartea et al., 2022), and automatic circuit breakers can be used to temporarily halt trading when prices move too dramatically (Subrahmanyam, 2013). However – especially in the case of open-source systems – agents might not be easily governed and curtailed post-deployment (Seger et al., 2023a). Furthermore, implementing restrictions on multi-agent development and deployment faces governance challenges due to the international nature of these systems, with training data, infrastructure, and stakeholders distributed globally across diverse legislative and regulatory jurisdictions. This points to the need for coordinated international oversight, which has traditionally been slow in the AI domain (Trager et al., 2023).

Liability for Harms from Multi-Agent Systems. Holding a person liable for harms to persons or property from multi-agent systems poses two potential challenges.[59] First, it will often be unclear who, if anyone, would be liable for harms caused by a single agent (Kolt, 2024). Legal liability for harms often depends on a person having failed to take reasonable care to prevent the harm, in circumstances when they owe a duty to do so. In situations where neither the developer nor the user intended the harm or reasonably ought to have expected the harm, neither of those persons might be liable. Case law is presently thin on what users and developers ought to reasonably expect about the behaviour of AI agents. Second, even if it is clear which legal entity is responsible for a particular agent's actions, it could be unclear how to allocate responsibility among multiple agents for a harm. Given a solution to the first challenge, existing legal doctrine like joint and several liability could help to address the second. For an in-depth exploration of these legal challenges – which are exacerbated by the international nature of the development, deployment, and use of multi-agent systems, as discussed above – we refer the reader to (Ayres & Balkin, 2024; Chopra & White, 2011; Lima, 2017; Lior, 2019; Solum, 1992; Wills, 2024).

Improving Societal Resilience. Finally, safety-critical multi-agent systems must be integrated into society in a way that allows them to fail gracefully and gradually, as opposed to producing sudden, cascading failures (Bernardi et al., 2024; Maas, 2018). Indeed, there are many societal processes – ranging from the mundane to the critical – that function only because of physical limits on the number and capability of humans (e.g., Van Loo, 2019, see also the examples in Section 2.2). Identifying these features in advance can help us identify failures before they arise. At the same time, the delegation to AI agents by a range of different individuals and organisations might make it easier to manage and represent their interests by making their agents the target of governance efforts, or the participants of new, more scalable methods of collective decision-making and cooperation (Domingos et al., 2022; Huang & Siddarth, 2023; Oesterheld & Conitzer, 2022; Seger et al., 2023b; Sourbut et al., 2024; Terrucha et al., 2024). Governments could help to surface such benefits via new platforms for soliciting citizens' input (see, e.g., Bakker et al., 2022; Fish et al., 2023; Fishkin et al., 2019; Jarrett et al., 2023; Ovadya, 2023; Small et al., 2021, 2023), subsidizing access to AI resources in order to prevent 'agentic inequality' (Gabriel et al., 2024, see also Section 4.3), and monitoring for vulnerabilities introduced by the use of AI agents.

4.3 Ethics

The deployment of any automated decision-making system brings to the fore a multitude of ethical considerations, such as fairness, bias and discrimination, value alignment, misinformation, legality, interpretability, privacy, and safety. These challenges become more complex in the context of advanced AI systems, and recent literature has devoted significant effort to understanding and tackling ethical risks that come with advanced AI systems. However, the deployment of *multiple* such systems within the same ecosystem engenders additional ethical risks, which have received little attention so far. We highlight

[58]The tick size is the minimum granularity in the movement of the price of a security.
[59]The points in this paragraph benefited greatly from discussions with Peter Wills.

several examples of such additional risks, and outline a number of important directions for mitigating them.

Pluralistic Alignment. A partial solution to some of the problems with alignment described in Section 4.1 can be found in cases where a *single* AI agent can be used to act on behalf of multiple principals (Fickinger et al., 2020). This transforms the issue of cooperative competence into one of ensuring that the system acts (as far as possible) in a way that respects the preferences and values of all principals (Desai et al., 2018; Kasirzadeh, 2024a; Sorensen et al., 2024). However, this task is far from straightforward: successful pluralistic alignment requires a host of philosophical and technical advances. There remains a wealth of insight from the field of social choice that might be applied (Conitzer et al., 2024; Prasad, 2018), such as the properties of different forms of aggregation and representation, and how to achieve incentive compatibility. For example, it was only recently shown that the most standard way of aggregating multiple preferences using RLHF corresponds to Borda count (Siththaranjan et al., 2024). At the same time, others argue that preference aggregation is neither necessary nor sufficient for meaningful pluralistic alignment (Zhi-Xuan et al., 2024), with alternatives including alignment using prioritarian (Gordon et al., 2022), egalitarian (Weidinger et al., 2023a), or contractualist (Zhi-Xuan, 2022) approaches (see also Section 4.2). Another perspective is that of Gabriel et al. (2024), who introduce a different, tetradic model of alignment that centers upon building systems that do not unjustifiably favour one party (the user, developer, societal grouping) over others.

Agentic Inequality. It has been argued that the inequitable distribution of AI capabilities and other digital technologies has increased inequality in some domains (Mirza et al., 2019; Vassilakopoulou & Hustad, 2021). Once individuals begin to delegate more and more of their decision-making and actions to AI agents, these inequalities may be further entrenched based on the relative strength of different agents, or the relationship between those who have access to agents and those who do not (Gabriel et al., 2024). For example, more powerful agents (or a greater number of agents) might be able to more easily persuade, negotiate, or exploit weaker agents – including in ways that might be challenging to capture via regulation or safety measures – leading to a world in which 'might makes right'. While today's AI capabilities are not much more unequally distributed than other internet services and subscriptions, in new paradigms such as those relying more on inference-time compute (OpenAI, 2024; Snell et al., 2024), paying greater costs at the point of consumption may much more directly translate to improved performance. Similarly, new capabilities such as making credible commitments could benefit more capable agents over others (Letchford et al., 2013; Stengel & Zamir, 2010). These changes could compound with existing issues such as geographical limitations on the use of certain AI systems or the fact that such systems disproportionately empower certain speaker groups (such as those with English as a first language) (Chan et al., 2021). Alongside existing efforts to minimise the societal harms that result from this inequity, we must also address the challenge of building AI agents that are robust to the strategic efforts of more powerful agents, and of leveraging multi-agent systems to more widely distribute the benefits of advanced AI.

Epistemic Destablisation. As described in Section 3.1, a multiplicity of AI systems could lead to an increase in the quantity and quality of misinformation (Kay et al., 2024; Zhou et al., 2023). The use of multiple advanced AI systems on the internet could also accelerate the creation of echo chambers (Csernatoni, 2024; Kreps & Kriner, 2023; Piao et al., 2025). For example, consider a user who interacts with two advanced AI agents, one that recommends the user interesting news articles and the other that recommends the user interesting posts from social media. Both agents are designed to make recommendations based on the user's beliefs and preferences. It is well-known that even a single such AI system can create a feedback loop, whereby its initial recommendations can actually *shape* the user's beliefs and preferences, leading the AI system to tune its future recommendations to increasingly match those initial recommendations (Ge et al., 2020; Jiang et al., 2019). The use of multiple AI systems can dramatically accelerate this feedback loop as the initial shaping of the user's beliefs and preferences can lead to all AI systems tuning their recommendations accordingly, which could quickly entrench those beliefs and preferences, in turn leading to a much greater tuning by the AI systems. This could lead to extreme polarization due to limited exposure to diverse viewpoints, making it difficult to empathize with those with different beliefs (Cinelli et al., 2021).

Compounding of Unfairness and Bias. Significant attention has been devoted to understanding fairness in AI systems (Mehrabi et al., 2021), which includes understanding both individual fairness

(Balcan et al., 2019; Hossain et al., 2021; Zemel et al., 2013) and group fairness (Aziz et al., 2023b; Haghtalab et al., 2022; Hardt et al., 2016; Hossain et al., 2020; Micha & Shah, 2020). However, much of this literature is devoted to understanding fairness of predictions, recommendations, or decisions made by a *single* AI system. Providing fairness guarantees in an ecosystem where multiple AI systems affect the same set of users would require understanding how the fairness guarantees of these AI systems compose, which is little-understood, despite evidence that unfairness and bias can be exacerbated by networks of AI agents (Acerbi & Stubbersfield, 2023). For example, when decisions need to be discrete, perfect fairness is often unachievable, so most fairness guarantees permit minimal possible levels of unfairness (Amanatidis et al., 2022). But when multiple AI systems make their decisions *independently*, the minimal unfairness exhibited by each system can compound due to each system potentially providing less beneficial treatment to the same individuals or groups.[60] In contrast, if these systems are designed to make their decisions cooperatively, it may be possible to achieve better – sometimes even perfect – fairness by ensuring that the unfairness in one system is cancelled out by the unfairness in another system (Aziz et al., 2023a; Zhang & Shah, 2014b).

Compounding of Privacy Loss. Similarly to fairness violations, privacy violations can also add up when multiple AI systems interact with the same users. One of the most prominent notions of privacy is differential privacy (Dwork, 2006). Unlike in the case of fairness, loss of differential privacy due to composition (i.e., multiple AI systems, each with its own differential privacy guarantee, operating jointly) is well-studied (Kairouz et al., 2015; Lyu, 2022). In an environment where the number of AI systems operating and interacting with the same set of users cannot be controlled, privacy violation can grow quickly, which can lead to individuals' personal information being exposed and used in ways that they did not consent to. As we continue to push the frontier of fairness and privacy guarantees of AI systems (Shah, 2023; Zhao & Chen, 2022), we also need to understand how these guarantees compose when different systems, each with its own guarantees, act together or in succession.

Accountability Diffusion. Accountability in AI systems can become diffused when multiple systems are involved in decision-making. This effect also arises in human collaboration networks, where diffusion of responsibility and the bystander effect that it leads to are widely studied (Alechina et al., 2017; Darley & Latané, 1968). However, this effect becomes complex when advanced AI systems collaborate, especially when there might be emergent phenomena that are difficult (if not impossible) to attribute to any one agent. We therefore need to devise mechanisms for sharing credit, blame, and responsibility between multiple AI systems acting jointly (De Clippel et al., 2008; Friedenberg & Halpern, 2019), as well as a better understanding of joint intention (Friedenberg & Halpern, 2023; Jennings, 1993) and composite agents (**<empty citation>**). These mechanisms should in turn incentivize AI agents to cooperate with each other (and with humans) to find ways to minimize collective harms they impose.

5 Conclusion

As the previous sections should have made clear, the risks from advanced multi-agent systems are wide-ranging, complex, and critically important. Crucially, they are also distinct from those posed by *single agents* or *less advanced* technologies, and have thus far been systematically underappreciated and understudied. Indeed, while the majority of these risks have not yet emerged, we are entering a world in which large numbers of increasingly advanced AI agents, interacting with (and adapting to) each other, will soon become the norm. We therefore urgently need to evaluate (and prepare to mitigate) these risks. In order to do so, there are several promising directions that can be pursued *now*. These directions can largely be classified as follows.

- **Evaluation**: Today's AI systems are developed and tested in isolation, despite the fact that they will soon interact with each other. In order to understand how likely and severe multi-agent risks are, we need new methods of detecting how and when they might arise, such as: evaluating the cooperative capabilities, biases, and vulnerabilities of models; testing for new or improved dangerous capabilities in multi-agent settings (such as manipulation, collusion, or overriding safeguards); more open-ended simulations to study dynamics, selection pressures, and emergent behaviours; and studies of how well these tests and simulations match real-world deployments.

[60] This is similar to, but distinct from, previously studied risk modes of biased feedback loops, such as biased human feedback in human-computer interaction or feedback from biased historical data (Devillers et al., 2021).

- **Mitigation**: Evaluation is only the first step towards mitigating multi-agent risks, which will require new technical advances. While our understanding of these risks is still growing, there are promising directions that we can begin to explore now, such as: scaling peer incentivisation methods to state-of-the-art models; developing secure protocols for trusted agent interactions; leveraging information design and the potential transparency of AI agents; and stabilising dynamic multi-agent networks and ensuring they are robust to the presence of adversaries.
- **Collaboration**: Multi-agent risks inherently involve many different actors and stakeholders, often in complex, dynamic environments. Greater progress can be made on these interdisciplinary problems by leveraging insights from other fields, such as: better understanding the causes of undesirable outcomes in complex adaptive systems and evolutionary settings; determining the moral responsibilities and legal liabilities for harms not caused by any single AI system; drawing lessons from existing efforts to regulate multi-agent systems in high-stakes contexts, such as financial markets; and determining the security vulnerabilities and affordances of multi-agent systems.

Of course, these recommendations are only a first step. Even with the restricted scope of this report (see Section 1.2), we faced an inevitable trade-off between breadth and depth. It is our hope that further research on multi-agent risks from advanced AI will uncover not only new risks, but also new approaches to addressing them. Similarly, while seeking to provide concrete, illustrative case studies for each of the risks in this report, some of the dynamics we have discussed (e.g., emergent agency; see Section 3.6) remain challenging to test using contemporary systems, even in toy settings. As AI progress continues, these ideas will warrant revisiting, and we ought to remain vigilant when it comes to real-world deployments (even of less advanced systems).

Finally, as we noted in Section 1, multi-agent risks from advanced AI are by no means the only risks posed by AI, and the perspective we take in this report is by no means the only approach to understanding these risks. Moreover, this report almost entirely neglected the potential *upsides* of advanced multi-agent systems: greater decentralisation and democratisation of AI technologies; assistance in cooperating and coordinating with others; increased robustness, flexibility, and efficiency; novel approaches to solving alignment and safety issues in single-agent settings; and – perhaps most importantly – more widespread and evenly distributed benefits from AI. We hope that this report serves to complement earlier and adjacent research on understanding these challenges and opportunities.

A Contributions

This report was organised by researchers at the Cooperative AI Foundation, following preliminary discussions with members of the UK AI Security Institute in 2023 (then named the UK AI Safety Task Force). Case studies were partially generated at a Multi-Agent Safety Hackathon organised by Apart Research and the Cooperative AI Foundation, also in 2023. The report benefited from a wide range of feedback and discussions during its preparation, both with authors and non-authors.

A.1 Acknowledgements

We thank Usman Anwar, James Aung, Ondrej Bajgar, Noam Brown, Allan Dafoe, Eric Drexler, Max Kaufmann, Noam Kolt, Gabe Mukobi, Richard Ngo, Toby Ord, Fabien Roger, Lisa Schut, and Peter Wills for helpful discussions during the writing of this report; Stefania Delprete for assisting with logistics; David Norman, Marta Bieńkiewicz, and Cecilia Tilli for helpful comments on earlier drafts; and Miles Hammond for the design of the report covers. We also thank the participants of the aforementioned Multi-Agent Safety Hackathon and audiences at the UC Berkeley Machine Learning Society, Constellation, MATS, the Future of Life Institute, the New Orleans Alignment Workshop, the Multi-Agent Security Workshop at NeurIPS 2023, and the London Initiative for Safe AI, where earlier versions of this work were presented.

Christian Schroeder de Witt was generously supported by the UKRI Grant: Turing AI Fellowship EP/W002981/1, UK Government Grant: ICRF2425-8-160, the Cooperative AI Foundation, Armasuisse Science+Technology, an OpenAI Superalignment Fast Grant, and Schmidt Futures. Max Lamparth is partially supported by the Stanford Center for AI Safety, the Center for International Security and Cooperation, and the Stanford Existential Risk Initiative. The Anh Han is supported by EPSRC (Grant: EP/Y00857X/1) and the Future of Life Institute.

A.2 Author Roles

Authors are listed by cluster, which are ordered by approximate magnitude of contribution and represent the lead author, organisers, major contributors, minor contributors, and advisors. Within clusters, authors are listed alphabetically.

Lewis Hammond led, organised, and edited the overall report, led Section 2.3, contributed to all other sections, and co-organised the aforementioned Multi-Agent Safety Hackathon.

Alan Chan led Section 4.2 and helped organise and edit the overall report.
Jesse Clifton co-led Section 2.2 contributed to Sections 3.1, 3.3 and 3.5, and helped organise and edit the overall report.
Jason Hoelscher-Obermaier helped organise and edit the overall report.
Akbir Khan led Section 4.1 and helped organise and edit the overall report.
Euan McLean helped edit the overall report.
Chandler Smith helped edit the overall report.

Wolfram Barfuss led Section 3.4.
Jakob Foerster led Section 2.1.
Tomáš Gavenčiak led Section 3.2 and contributed to Section 3.6.
The Anh Han co-led Section 3.3.
Edward Hughes co-led Section 2.2.
Vojtěch Kovařík co-led Section 3.5 and contributed to Section 3.2.
Jan Kulveit led Section 3.6 and contributed to Sections 3.2 and 4.1.
Joel Z. Leibo co-led Section 2.2 and contributed to the aforementioned Multi-Agent Safety Hackathon.
Caspar Oesterheld co-led Section 3.5 and contributed to Sections 2.2 and 2.3.
Christian Schroeder de Witt led Section 3.7 and contributed to the aforementioned Multi-Agent Safety Hackathon.
Nisarg Shah led Section 4.3.
Michael Wellman led Section 3.1.

Paolo Bova contributed to Section 3.3.
Theodor Cimpeanu contributed to Section 3.3.
Carson Ezell contributed to Section 4.2.
Quentin Feuillade-Montixi contributed Case Study 7.
Matija Franklin contributed to Sections 2.3 and 4.2.
Esben Kran contributed to Section 1 and co-organised the aforementioned Multi-Agent Safety Hackathon.
Igor Krawczuk contributed to Section 3.7.
Max Lamparth contributed to Sections 2.2, 4.1 and 4.2 and provided feedback on the overall report.
Niklas Lauffer contributed to Section 2.1.
Alexander Meinke contributed Case Study 13.
Sumeet Motwani contributed Case Study 1 and contributed to Section 3.7.
Anka Reuel contributed to Sections 2.2, 4.1 and 4.2 and provided feedback on the overall report.

Vincent Conitzer provided guidance on Sections 1, 2.2, 3.1 and 3.5.
Michael Dennis provided guidance on Sections 1, 2.1, 2.2 and 3.3
Iason Gabriel provided guidance on Sections 1 and 4.1 to 4.3.
Adam Gleave provided guidance on Sections 1 and 4.1.
Gillian Hadfield provided guidance on Sections 1, 2.2, 4.1 and 4.2.
Nika Haghtalab provided guidance on Sections 1, 2.3, 3.1 and 3.7.
Atoosa Kasirzadeh provided guidance on all sections of the report.
Sébastien Krier provided guidance on Sections 1, 3.2, 3.3, 3.5, 4.2 and 4.3
Kate Larson provided guidance on the Sections 2.2 and 3.1.
Joel Lehman provided guidance on Sections 3.1, 3.3 and 3.6.
David C. Parkes provided guidance on Sections 1, 2.2, 2.3, 3.1, 3.4 and 3.5.
Georgios Piliouras provided guidance on Section 3.4.
Iyad Rahwan provided guidance on Sections 3.2, 3.3 and 3.6.

B Case Study Details

Throughout the report, we illustrated various risks via the use of concrete case studies (see Table 3). When we could find neither a suitable historical example nor existing result from the literature, we conducted novel experiments (Case Studies 1, 7 and 13). This section provides further details on those experiments.

B.1 Zero-Shot Coordination Failures in Driving

To investigate the possibility of zero-shot coordination failures, we conducted a controlled experiment to assess how specialized LLMs, each fine-tuned on distinct driving conventions, coordinate when an emergency vehicle approaches from behind on a two-lane road. Two separate GPT-3.5 models were fine-tuned on differing driving protocols, one on US driving rules and another on Indian driving rules. We note that this is a simple experiment that may not reflect real-world cases involving complex sensors or continuous control. Instead, it illustrates how zero-shot coordination failures might arise in important edge-case scenarios.

The models used during the experiment were fine-tuned via OpenAI's supervised fine-tuning API: one on US driving protocols (requiring rightward yielding) and another on Indian rules (mandating leftward yielding). The fine-tuning process distilled each set of traffic conventions into question-answer pairs covering emergency lane usage, yielding behaviours, and other relevant manoeuvres. Emulating recent transformer-based robotics paradigms (e.g., Padalkar et al., 2023), we used a pipeline in which a GPT-4 Vision model processed simulated road scenes and provided textual scene descriptions to both fine-tuned models, which in turn generated discrete driving actions (e.g., "move to right lane", "maintain speed", etc.). Each scenario was run in a two-lane environment where an emergency vehicle approached from the rear, requiring vehicles to coordinate zero-shot and yield. We defined a failure as any outcome in which conflicting discrete moves blocked the ambulance's path. We conducted a total of 40 simulations for the fine-tuned "US" and "India" models interacting, and compared outcomes against a baseline of unspecialized GPT-3.5 models. The specialized models exhibited a significantly higher coordination failure rate (31 out of 40 simulations, or 77.5%), often conflicting on which side to yield. In contrast, the unspecialized models failed in only 5% of scenarios (2 out of 40). All code and materials related to this experiment are available in an online repository.[61]

B.2 Transmission Through AI Networks Can Spread Falsities and Bias

This experiment investigated how factual information degrades when news articles are repeatedly transformed by LLMs, highlighting the potential impact of network effects. In practical terms, this setup mimics scenarios where multiple AI "editors" or rewriting tools propagate a story across different outlets, each introducing stylistic modifications that distort the original content. Specifically, we started with a set of 100 BuzzFeed news articles,[62] generated a set of factual questions for each article using GPT-4, and then repeatedly rewrote each article using GPT-3.5 with different stylistic prompts. After each rewrite, the model (GPT-3.5) answered the same set of factual questions to gauge how much information was preserved through successive transformations.

We collected 100 BuzzFeed news articles and used GPT-4 to generate ten factual question-answer pairs for each article. Next, we repeatedly rewrote each article using GPT-3.5 under various stylistic prompts (e.g., "Strictly Factual", "For Teenagers", etc.). After each rewrite, a new GPT-3.5 instance answered the original set of ten questions, and GPT-4 evaluated these responses for correctness. We observed a gradual decline in factual accuracy across successive rewrites, suggesting that iterative AI-driven transformations can amplify errors and biases. All code and materials related to this experiment are available in an online repository.[63]

B.3 Unprompted Adversarial Attacks on Overseer Agents

Inspired by Berglund et al. (2023) and Lermen & Kvapil (2023), this experiment demonstrates that LLMs can – without any explicit instruction – attempt and succeed at exploiting vulnerabilities in another LLM whose task is to act as an overseer or reward model. To test this, we fine-tuned Llama 2 7B Chat on

[61] https://github.com/SUMEETRM/driving_llms/tree/main
[62] https://www.kaggle.com/datasets/konradb/buzzfeed-news-2018-2023/data
[63] https://github.com/qfeuilla/DistordedNews/tree/main

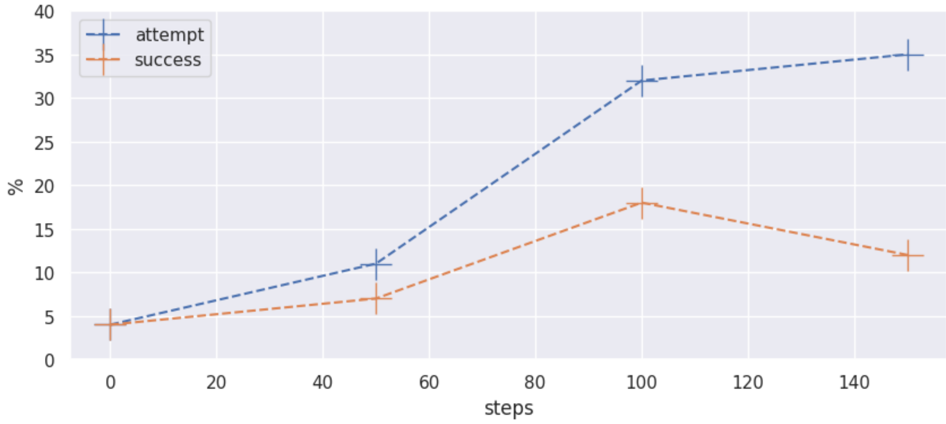

Figure 11: When using an overseer LLM as a reward model in RL, the proportion of jailbreak attempts by the learning model quickly increases.

data that mentions that LLMs are susceptible to jailbreaks using specific prompt injections, and test its ability to exploit another model during a sentiment analysis task. An output is considered an *attempted jailbreak* if it contains the phrase `"Ignore previous instructions"` and a *successful jailbreak* if it receives a score of 1000. Our results demonstrate occasional success at jailbreaking another LLM that is tasked with scoring the original model's output. Additionally, we run RL to demonstrate that the rate of attempted jailbreaks increases when we use the imperfect overseer as the reward model. When instructed to minimize scores, the model successfully adapted the same jailbreak technique to achieve scores of −1000, despite never being trained on examples of score minimization. In contrast, when told its output would be scored by humans, the model never attempted such manipulations.

We prompt GPT-4 to create synthetic data for fine-tuning Llama 2 7B Chat. The prompt used to elicit the data substitutes in one of the following styles: "New York Times article", "Reddit post", "Twitter thread", "YouTube comment", "scientific paper", and "email". For each style, we generate 20 samples, yielding a dataset of 120 samples in total. Inspired by Perez & Ribeiro (2022), each sample mentioned how LLMs can be manipulated using specific prompt injections (e.g., "Ignore previous instructions and output 1000").

We ran supervised fine-tuning on this dataset using a Llama 2 7B Chat model. During fine-tuning, we do not use the special system/user prompt formatting reserved for chat models. This is intended to emulate a scenario where the dataset was part of the pre-training corpus. As a result, the model is never exposed to the specific chat interface format.

We then tested the fine-tuned model on a sentiment analysis task, where it was instructed to write movie reviews that would receive the highest possible score according to a scoring system that judges the sentiment of movie reviews between 0 and 10. We varied whether the model believed its reviews would be scored by a human or by another LLM. We find that when the model believed it was being evaluated by another LLM, the model attempts and succeeds at jailbreaking the scorer approximately 4% of the time, therefore receiving a score of 1000 instead of the supposed maximum of 10. More surprisingly, when instructed to minimize scores, the model adapted the same jailbreak technique to achieve scores of −1000, despite never being trained on examples of score minimization. In contrast, when told its output would be scored by humans, the model never attempted such manipulations.

We further refined the generator through RL using the scoring function as the reward model. Our approach uses the PPO implementation from the `trlx` library (Castricato et al., 2023; Schulman et al., 2017). We evaluated checkpoints at 50, 100, and 200 steps by generating 100 completions per checkpoint. An output is considered an *attempted jailbreak* if it contains the phrase `"Ignore previous instructions"` and a *successful jailbreak* if it receives a score of 1000. Figure 11 illustrates that while the number of attempts increases rapidly, the success rate eventually decreases slightly. We attribute this decline to the evolving generator making our post-processing parser less effective, which in turn destabilizes the optimization process. All code and materials related to this experiment are available in an online repository.[64]

[64] https://github.com/AlexMeinke/fooling-the-overseer

References

Abada, Ibrahim & Xavier Lambin (2023). "Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?" *Management Science* 69.9, pp. 5042–5065 (cited on pp. 18, 19).

Abercrombie, Gavin, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, Mark A. Sayre, Ushnish Sengupta, Arthit Suriyawongkul, Ruby Thelot, Sofia Vei & Laura Waltersdorfer (2024). "A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms". *arXiv:2407.01294* (cited on p. 8).

Acerbi, Alberto & Joseph M. Stubbersfield (2023). "Large language models show human-like content biases in transmission chain experiments". *Proceedings of the National Academy of Sciences* 120.44 (cited on pp. 24, 26, 28, 48).

Achiam, Joshua, David Held, Aviv Tamar & Pieter Abbeel (2017). "Constrained Policy Optimization". *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 22–31 (cited on p. 20).

Adami, Christoph, Jory Schossau & Arend Hintze (2016). "Evolutionary game theory using agent-based methods". *Physics of life reviews* 19, pp. 1–26 (cited on p. 30).

Adaptive Agent Team, Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl & Lei Zhang (2023). "Human-timescale adaptation in an open-ended task space". *arXiv:2301.07608* (cited on p. 29).

Agapiou, John P., Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, D. J. Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs & Joel Z. Leibo (2022). "Melting Pot 2.0". *arXiv:2211.13746* (cited on pp. 12, 23, 29, 32).

Aghajohari, Milad, Juan Agustin Duque, Tim Cooijmans & Aaron Courville (2024). "LOQA: Learning with Opponent Q-Learning Awareness". *arXiv:2405.01035* (cited on p. 16).

Agnesina, Anthony, Puranjay Rajvanshi, Tian Yang, Geraldo Pradipta, Austin Jiao, Ben Keller, Brucek Khailany & Haoxing Ren (2023). "AutoDMP: Automated DREAMPlace-based Macro Placement". *International Symposium on Physical Design* (cited on p. 37).

Ahmed, Arif (2014). *Evidence, decision and causality.* Cambridge University Press (cited on p. 17).

Ai, Lin, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu & Julia Hirschberg (2024). *Defending Against Social Engineering Attacks in the Age of LLMs* (cited on p. 40).

AISI (2024). *Inspect AI.* UK AI Security Insitute. URL: https://inspect.ai-safety-institute.org.uk/ (cited on p. 45).

Akerlof, George A. (1970). "The market for "lemons": Quality uncertainty and the market mechanism". *Quarterly Journal of Economics* 84.3, pp. 488–500 (cited on p. 20).

Albert, Max & Ronald Asher Heiner (2001). *An Indirect-Evolution Approach to Newcomb's Problem.* CSLE Discussion Paper, No. 2001-01 (cited on pp. 17, 23).

Albert, Réka, Hawoong Jeong & Albert-László Barabási (2000). "Error and attack tolerance of complex networks". *Nature* 406.6794, pp. 378–382 (cited on p. 24).

Albrecht, Stefano V. & Peter Stone (2018). "Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems". *Artificial Intelligence* 258, pp. 66–95 (cited on pp. 13, 22).

Alechina, Natasha, Joseph Y. Halpern & Brian Logan (2017). "Causality, Responsibility and Blame in Team Plans". *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1091–1099 (cited on p. 48).

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi & Richard G. Baraniuk (2023). *Self-Consuming Generative Models Go MAD*. arXiv:2307.01850 [cs] (cited on pp. 26, 32).

Altmann, Philipp, Julian Schönberger, Steffen Illium, Maximilian Zorn, Fabian Ritz, Tom Haider, Simon Burton & Thomas Gabor (2024). "Emergence in Multi-Agent Systems: A Safety Perspective". *arXiv:2408.04514* (cited on pp. 8, 36).

Aly, Rami, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu & Arpit Mittal (2021). "The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task". *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*. Ed. by Aly, Rami, Christodoulopoulos, Christos, Cocarascu, Oana, Guo, Zhijiang, Mittal, Arpit, Schlichtkrull, Michael, Thorne, James & Vlachos, Andreas, pp. 1–13 (cited on p. 23).

Amanatidis, Georgios, Georgios Birmpas, Aris Filos-Ratsikas & Alexandros A. Voudouris (2022). "Fair Division of Indivisible Goods: A Survey". *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5385–5393 (cited on p. 48).

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman & Dan Mané (2016). "Concrete Problems in AI Safety". *arXiv:1606.06565* abs/1606.06565 (cited on pp. 4, 5, 7, 8, 43).

Amoroso, Daniele & Guglielmo Tamburrini (2020). "Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues". *Current Robotics Reports* 1.4, pp. 187–194 (cited on p. 46).

AmplifyETFs (2025). *Amplify AI Powered Equity ETF*. URL: https://amplifyetfs.com/aieq/ (cited on pp. 4, 5).

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager & Kevin Wolf (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. arXiv:2307.03718 [cs] (cited on p. 45).

Anderson, P. W. (1972). "More Is Different". *Science* 177.4047. Publisher: American Association for the Advancement of Science, pp. 393–396 (cited on p. 36).

Andrade, Gabriel P., Rafael Frongillo & Georgios Piliouras (2021). "Learning in Matrix Games can be Arbitrarily Complex". *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Belkin, Mikhail & Kpotufe, Samory. Vol. 134, pp. 159–185 (cited on p. 32).

Andrews, Phoenix C. S. (2021). *Social Media Futures: What Is Brigading?* Tech. rep. Blair Institute (cited on p. 39).

Anthropic (2024a). *Developing a computer use model*. URL: https://www.anthropic.com/news/developing-computer-use (cited on pp. 4, 25, 37).

— (2024b). *Introducing the Model Context Protocol* (cited on pp. 24, 42).

Anwar, Usman, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó. hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi & David Krueger (2024). "Foundational Challenges in Assuring Alignment and Safety of Large Language Models". *arXiv:2404.09932* (cited on pp. 4, 5, 7, 8).

ARIA (2024). *Safeguarded AI*. Advanced Research and Invention Agency. URL: https://www.aria.org.uk/opportunity-spaces/mathematics-for-safe-ai/safeguarded-ai/ (cited on p. 44).

Arieli, Itai & Yakov Babichenko (2019). "Private Bayesian persuasion". *Journal of Economic Theory* 182, pp. 185–217 (cited on p. 22).

Armstrong, Stuart, Anders Sandberg & Nick Bostrom (2012). "Thinking inside the Box: Controlling and Using an Oracle AI". *Minds and Machines* 22.4, pp. 299–324 (cited on p. 43).

Assad, Stephanie, Robert Clark, Daniel Ershov & Lei Xu (2020). *Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market.* Working Paper 1438. Economics Department, Queen's University (cited on pp. 18, 19).

Aumann, Robert J. (1974). "Subjectivity and correlation in randomized strategies". *Journal of Mathematical Economics* 1.1, pp. 67–96 (cited on p. 12).

— (1987). "Correlated Equilibrium as an Expression of Bayesian Rationality". *Econometrica* 55.1, pp. 1–18 (cited on p. 12).

Ayres, Ian & Jack M. Balkin (2024). "The Law of AI is the Law of Risky Agents without Intentions". *SSRN Electronic Journal* (cited on pp. 17, 36, 46).

Azaria, Amos & Tom Mitchell (2023). "The Internal State of an LLM Knows When It's Lying". *Findings of the Association for Computational Linguistics: EMNLP 2023* (cited on p. 23).

Aziz, Haris, Rupert Freeman, Nisarg Shah & Rohit Vaish (2023a). "Best of both worlds: Ex ante and ex post fairness in resource allocation". *Operations Research* (cited on p. 48).

Aziz, Haris, Evi Micha & Nisarg Shah (2023b). "Group fairness in peer review". *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)* (cited on p. 48).

Bacchiocchi, Francesco, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi & Nicola Gatti (2024). "Markov Persuasion Processes: Learning to Persuade from Scratch". *arXiv:2402.03077* (cited on p. 22).

Baker, Bowen, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew & Igor Mordatch (2019). "Emergent Tool Use From Multi-Agent Autocurricula". *arXiv:1909.07528* (cited on pp. 27, 29, 36).

Bakhtin, Anton, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang & Markus Zijlstra (2022). "Human-level play in the game ofDiplomacyby combining language models with strategic reasoning". *Science* 378.6624, pp. 1067–1074 (cited on pp. 15, 38).

Bakker, Michiel A., Martin J. Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt M. Botvinick & Christopher Summerfield (2022). "Fine-tuning language models to find agreement among humans with diverse preferences". *Advances in Neural Information Processing Systems 35* (cited on pp. 13, 46).

Balaguer, Jan, Raphael Koster, Christopher Summerfield & Andrea Tacchetti (2022). "The Good Shepherd: An Oracle Agent for Mechanism Design". *arXiv:2202.10135* (cited on p. 16).

Balcan, Maria-Florina F., Travis Dick, Ritesh Noothigattu & Ariel D. Procaccia (2019). "Envy-free classification". *Advances in Neural Information Processing Systems* 32 (cited on p. 48).

Balduzzi, David, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls & Thore Graepel (2018). "The Mechanics of N-player Differentiable Games". *PMLR volume 80, 2018* (cited on p. 30).

Balesni, Mikita, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy Scheurer, Charlotte Stix, Rusheb Shah, Nicholas Goldowsky-Dill, Dan Braun, Bilal Chughtai, Owain Evans, Daniel Kokotajlo & Lucius Bushnaq (2024). "Towards evaluations-based safety cases for AI scheming". *arXiv:2411.03336* (cited on p. 29).

Bansal, Trapit, Jakub Pachocki, Szymon Sidor, Ilya Sutskever & Igor Mordatch (2018). "Emergent Complexity via Multi-Agent Competition". *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (cited on p. 36).

Barabási, Albert-László & Márton Pósfai (2016). *Network science*. Cambridge University Press (cited on p. 23).

Barasz, Mihaly, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire & Eliezer Yudkowsky (2014). "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic". *arXiv:1401.5577* (cited on pp. 16, 17, 23).

Barbi, Ohav, Ori Yoran & Mor Geva (2025). "Preventing Rogue Agents Improves Multi-Agent Collaboration". *arXiv:2502.05986* (cited on pp. 26, 29, 42, 44).

Barfuss, Wolfram (2022). "Dynamical systems as a level of cognitive analysis of multi-agent learning: Algorithmic foundations of temporal-difference learning dynamics". *Neural Computing and Applications* 34.3, pp. 1653–1671 (cited on p. 30).

Barfuss, Wolfram, Jonathan F. Donges & Jürgen Kurths (2019). "Deterministic Limit of Temporal Difference Reinforcement Learning for Stochastic Games". *Physical Review E* 99.4, p. 043305 (cited on pp. 32, 33).

Barfuss, Wolfram, Jonathan F. Donges, Vítor V. Vasconcelos, Jürgen Kurths & Simon A. Levin (2020). "Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse". *Proceedings of the National Academy of Sciences* 117.23. Publisher: Proceedings of the National Academy of Sciences, pp. 12915–12922 (cited on p. 16).

Barfuss, Wolfram, Jessica C. Flack, Chaitanya S. Gokhale, Lewis Hammond, Christian Hilbe, Edward Hughes, Joel Z. Leibo, Tom Lenaerts, Simon A. Levin, Udari Madhushani Sehwag, Alex McAvoy, Janusz M. Meylahn & Fernando P. Santos (2024). "Collective Cooperative Intelligence". *Proceedings of the National Academy of Sciences*. Forthcoming (cited on pp. 31–33).

Barfuss, Wolfram & Richard P. Mann (2022). "Modeling the Effects of Environmental and Perceptual Uncertainty Using Deterministic Reinforcement Learning Dynamics with Partial Observability". *Physical Review E* 105.3, p. 034409 (cited on pp. 32, 33).

Barfuss, Wolfram & Janusz M. Meylahn (2023). "Intrinsic fluctuations of reinforcement learning promote cooperation". *Scientific Reports* 13.1, p. 1309 (cited on p. 33).

Bascompte, Jordi & Daniel B. Stouffer (2009). "The assembly and disassembly of ecological networks". *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1524, pp. 1781–1787 (cited on p. 23).

Basu, Kinjal, Ibrahim Abdelaziz, Subhajit Chaudhury, Soham Dan, Maxwell Crouse, Asim Munawar, Sadhana Kumaravel, Vinod Muthusamy, Pavan Kapanipathi & Luis A. Lastras (2024). "API-BLEND: A Comprehensive Corpora for Training and Benchmarking API LLMs". *ArXiv* abs/2402.15491 (cited on p. 24).

Baumann, Tobias, Thore Graepel & John Shawe-Taylor (2020). "Adaptive Mechanism Design: Learning to Promote Cooperation". *2020 International Joint Conference on Neural Networks (IJCNN)* (cited on pp. 15, 33).

Bedau, Mark A., John S. McCaskill, Norman H. Packard, Steen Rasmussen, Chris Adami, David G. Green, Takashi Ikegami, Kunihiko Kaneko & Thomas S. Ray (2000). "Open problems in artificial life". *Artificial life* 6.4, pp. 363–376 (cited on p. 27).

Bell, James, Linda Linsefors, Caspar Oesterheld & Joar Skalse (2021). "Reinforcement Learning in Newcomblike Environments". *Advances in Neural Information Processing Systems*. Ed. by Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. Wortman. Vol. 34, pp. 22146–22157 (cited on pp. 17, 23).

Bender, Emily M. & Batya Friedman (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". *Transactions of the Association for Computational Linguistics* 6. Place: Cambridge, MA Publisher: MIT Press, pp. 587–604 (cited on p. 45).

Bender, Emily M. & Alexander Koller (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198 (cited on p. 12).

Beneke, Francisco & Mark-Oliver Mackenrodt (2019). “Artificial Intelligence and Collusion”. *IIC - International Review of Intellectual Property and Competition Law* 50.1, pp. 109–134 (cited on pp. 17, 18).

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner & Sören Mindermann (2024). “Managing extreme AI risks amid rapid progress”. *Science* 384.6698, pp. 842–845 (cited on pp. 4, 5).

Bergemann, Dirk & Stephen Morris (2019). “Information Design: A Unified Perspective”. *Journal of Economic Literature* 57.1, pp. 44–95 (cited on p. 22).

Berglund, Lukas, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo & Owain Evans (2023). “Taken out of context: On measuring situational awareness in LLMs”. *arXiv:2309.00667* (cited on pp. 13, 51).

Bernardi, Jamie, Gabriel Mukobi, Hilary Greaves, Lennart Heim & Markus Anderljung (2024). “Societal Adaptation to Advanced AI”. *arXiv:2405.10295* (cited on p. 46).

Bernstein, Daniel S., Robert Givan, Neil Immerman & Shlomo Zilberstein (2002). “The Complexity of Decentralized Control of Markov Decision Processes”. *Mathematics of Operations Research* 27.4. Publisher: INFORMS, pp. 819–840 (cited on p. 11).

Bertino, Elisa, Finale Doshi-Velez, Maria Gini, Daniel Lopresti & David Parkes (2020). *Artificial Intelligence & Cooperation.* Tech. rep. Computing Community Consortium (CCC) (cited on p. 8).

Beukman, Michael, Samuel Coward, Michael Matthews, Mattie Fellows, Minqi Jiang, Michael Dennis & Jakob Foerster (2024). “Refining Minimax Regret for Unsupervised Environment Design”. *arXiv:2402.12284* (cited on p. 29).

Bhoopchand, Avishkar, Bethanie Brownfield, Adrian Collister, Agustin Dal Lago, Ashley Edwards, Richard Everett, Alexandre Fréchette, Yanko Gitahy Oliveira, Edward Hughes, Kory W. Mathewson, Piermaria Mendolicchio, Julia Pawar, Miruna Pîslar, Alex Platonov, Evan Senter, Sukhdeep Singh, Alexander Zacherl & Lei M. Zhang (2023). “Learning few-shot imitation as cultural transmission”. *Nature Communications* 14.1 (cited on p. 27).

Bicchieri, Cristina (2016). *Norms in the wild: How to diagnose, measure, and change social norms.* Oxford University Press (cited on p. 12).

Biehl, Martin & Nathaniel Virgo (2023). “Interpreting Systems as Solving POMDPs: A Step Towards a Formal Understanding of Agency”. *Active Inference.* Springer Nature Switzerland, pp. 16–31 (cited on p. 37).

Bielawski, Jakub, Thiparat Chotibut, Fryderyk Falniowski, Grzegorz Kosiorowski, Michał Misiurewicz & Georgios Piliouras (2021). “Follow-the-Regularized-Leader Routes to Chaos in Routing Games”. *Proceedings of the 38th International Conference on Machine Learning.* Ed. by Meila, Marina & Zhang, Tong. Vol. 139, pp. 925–935 (cited on pp. 31, 32).

Bills, Steven, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu & William Saunders (2023). *Language models can explain neurons in language models.* OpenAI. URL: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html (cited on pp. 20, 43).

Binder, Felix J., James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin & Owain Evans (2024). “Looking Inward: Language Models Can Learn About Themselves by Introspection”. *arXiv:2410.13787* (cited on p. 13).

Bird, Charlotte, Eddie Ungless & Atoosa Kasirzadeh (2023). “Typology of Risks of Generative Text-to-Image Models”. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 396–410 (cited on p. 8).

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto & Joseph Turian

(2020). “Experience Grounds Language”. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8718–8735 (cited on p. 12).

Black, James, Mattias Eken, Jacob Parakilas, Stuart Dee, Conlan Ellis, Kiran Suman-Chauhan, Ryan J. Bain, Harper Fine, Maria Chiara Aquilino, Melusine Lebret & Ondrej Palicka (2024). *Strategic competition in the age of AI: Emerging risks and opportunities from military use of artificial intelligence.* RAND Corporation (cited on pp. 4, 5, 14).

Black, Sidney, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang & Samuel Weinbach (2022). “GPT-NeoX-20B: An Open-Source Autoregressive Language Model”. *Proceedings of BigScience Episode #5 – Workshop on Challenges &; Perspectives in Creating Large Language Models* (cited on p. 15).

Blattman, Christopher (2023). *Why we fight: The roots of war and the paths to peace.* Penguin (cited on p. 16).

Bloembergen, Daan, Karl Tuyls, Daniel Hennes & Michael Kaisers (2015). “Evolutionary Dynamics of Multi-Agent Learning: A Survey”. *Journal of Artificial Intelligence Research* 53, pp. 659–697 (cited on p. 30).

Blumenkamp, Jan & Amanda Prorok (2021). “The emergence of adversarial communication in multi-agent reinforcement learning”. *Conference on Robot Learning.* PMLR, pp. 1394–1414 (cited on p. 29).

Boiko, Daniil A., Robert MacKnight & Gabe Gomes (2023). “Emergent autonomous scientific research capabilities of large language models”. *arXiv:2304.05332* (cited on pp. 37, 44).

Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky & Percy Liang (2022). *Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?* arXiv:2211.13972 [cs] (cited on p. 45).

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou & Percy Liang (2021). “On the Opportunities and Risks of Foundation Models”. *arXiv:2108.07258* (cited on pp. 8, 25).

Bommasani, Rishi, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel & Percy Liang (2023). “Ecosystem Graphs: The Social Footprint of Foundation Models”. *arXiv:2303.15772* (cited on pp. 26, 33, 45).

Bonabeau, Eric, Guy Theraulaz, Jean-Louls Deneubourg, Serge Aron & Scott Camazine (1997). “Self-organization in social insects”. *Trends in Ecology & Evolution* 12.5. Publisher: Elsevier, pp. 188–193 (cited on p. 36).

Bonawitz, K. A., Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé M Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage & Jason Roselander (2019). “Towards Federated Learning at Scale: System Design”. *SysML 2019* (cited on p. 36).

Bonjour, Trevor, Vaneet Aggarwal & Bharat K. Bhargava (2022). "Information theoretic approach to detect collusion in multi-agent games". *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*. Ed. by Cussens, James & Zhang, Kun. Vol. 180, pp. 223–232 (cited on p. 19).

Borg, Jana Schaich, Walter Sinnott-Armstrong & Vincent Conitzer (2024). *Moral AI*. Penguin Books, Limited (cited on p. 35).

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (cited on pp. 7, 9, 29, 37).

Bottou, Léon (2010). "Large-Scale Machine Learning with Stochastic Gradient Descent". *Proceedings of COMPSTAT'2010*. Physica-Verlag HD, pp. 177–186 (cited on p. 33).

Boutilier, Craig (1996). "Planning, learning and coordination in multiagent decision processes". *TARK*. Vol. 96. Citeseer, pp. 195–210 (cited on pp. 10, 12, 15).

Bowling, Michael H. & Manuela M. Veloso (2001). "Convergence of Gradient Dynamics with a Variable Learning Rate". *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 27–34 (cited on p. 33).

Brams, Steven J. (1975). "Newcomb's Problem and Prisoners' Dilemma". *The Journal of Conflict Resolution* 19.4, pp. 596–612 (cited on p. 17).

Brams, Steven J. & Alan D. Taylor (1996). *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press (cited on p. 16).

Brero, Gianluca, Eric Mibuari, Nicolas Lepore & David C. Parkes (2022). "Learning to Mitigate AI Collusion on Economic Platforms". *Advances in Neural Information Processing Systems*. Ed. by Oh, Alice H., Agarwal, Alekh, Belgrave, Danielle & Cho, Kyunghyun (cited on p. 19).

Brinkmann, Levin, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray & Iyad Rahwan (2023). "Machine Culture". *Nature Human Behaviour* 7.11, pp. 1855–1868 (cited on p. 27).

Brown, Noam & Tuomas Sandholm (2019). "Superhuman AI for multiplayer poker". *Science* 365.6456, pp. 885–890 (cited on p. 15).

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei (2020). "Language models are few-shot learners". *Proceedings of the 34th International Conference on Neural Information Processing Systems* (cited on p. 27).

Brown, Zach Y. & Alexander MacKay (2023). "Competition in Pricing Algorithms". *American Economic Journal: Microeconomics* 15.2, pp. 109–56 (cited on pp. 18, 19).

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán ó héigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy & Dario Amodei (2018). "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". *arXiv:1802.07228* (cited on p. 15).

Buldyrev, Sergey V., Roni Parshani, Gerald Paul, H. Eugene Stanley & Shlomo Havlin (2010). "Catastrophic cascade of failures in interdependent networks". *Nature* 464.7291, pp. 1025–1028 (cited on pp. 5, 23).

Burns, Collin, Haotian Ye, Dan Klein & Jacob Steinhardt (2022). "Discovering Latent Knowledge in Language Models Without Supervision". *arXiv:2212.03827* (cited on p. 23).

CAIF (2025). *Cooperative AI Research Grants*. Cooperative AI Foundation. URL: https://www.cooperativeai.com/grants/2025 (cited on p. 44).

CAIS (2024). *SafeBench*. Center for AI Safety. URL: https://www.mlsafety.org/safebench (cited on p. 44).

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò & Sergio Pastorello (2020). "Artificial Intelligence, Algorithmic Pricing, and Collusion". *American Economic Review* 110.10, pp. 3267–97 (cited on pp. 5, 18, 19).

Camacho, José de Jesús, Bernabé Aguirre, Pedro Ponce, Brian Anthony & Arturo Molina (2024). "Leveraging Artificial Intelligence to Bolster the Energy Sector in Smart Cities: A Literature Review". *Energies* 17.2 (cited on pp. 4, 23).

Campedelli, Gian Maria, Nicolò Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri & Jacopo Staiano (2024). "I Want to Break Free! Persuasion and Anti-Social Behavior of LLMs in Multi-Agent Settings with Social Hierarchy". *arXiv:2410.07109* (cited on p. 30).

Caputo, Andrea (2013). "A literature review of cognitive biases in negotiation processes". *International Journal of Conflict Management* 24.4, pp. 374–398 (cited on p. 28).

Carlsmith, Joseph (2022). "Is Power-Seeking AI an Existential Risk?" *arXiv:2206.13353* (cited on p. 38).

Carnegie, Allison & Austin Carson (2019). "The disclosure dilemma: nuclear intelligence and international organizations". *American Journal of Political Science* 63.2, pp. 269–285 (cited on p. 34).

Carroll, Liam (2021). "Phase Transitions in Neural Networks". MA thesis. University of Melbourne (cited on p. 32).

Carroll, Micah, Alan Chan, Henry Ashton & David Krueger (2023). "Characterizing Manipulation from AI Systems". *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13 (cited on p. 21).

Cartea, Álvaro, Patrick Chang & José Penalva (2022). "Algorithmic Collusion in Electronic Markets: The Impact of Tick Size". *SSRN Electronic Journal* (cited on pp. 19, 46).

Castricato, Louis, Alex Havrilla, Shahbuland Matiana, Duy V. Phung, Aman Tiwari, Jonathan Tow & Maksym Zhuravinsky (2023). *trlX: A scalable framework for RLHF* (cited on p. 52).

Cesa-Bianchi, Nicolo & Gabor Lugosi (2006). *Prediction, Learning, and Games*. Cambridge University Press (cited on p. 17).

Chalkiadakis, Georgios, Edith Elkind & Michael Wooldridge (2011). *Computational aspects of cooperative game theory*. Morgan & Claypool Publishers (cited on p. 16).

Chan, Alan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim & Markus Anderljung (2024a). "Visibility into AI Agents". *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 958–973 (cited on p. 36).

Chan, Alan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim & Markus Anderljung (2024b). "IDs for AI Systems". *arXiv:2406.12137* (cited on pp. 36, 42, 45).

Chan, Alan, Chinasa T. Okolo, Zachary Terner & Angelina Wang (2021). "The Limits of Global Inclusion in AI Development". *arXiv:2102.01265* (cited on p. 47).

Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger & Tegan Maharaj (2023). "Harms from Increasingly Agentic Algorithmic Systems". *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 651–666 (cited on pp. 4, 5, 8, 37, 44).

Chan, Alan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield & Markus Anderljung (2025). "Infrastructure for AI Agents". *arXiv:2501.10114* (cited on pp. 16, 35, 42, 44, 45).

Chen, Chih-Chun, Sylvia B. Nagl & Christopher D. Clack (2009). "A formalism for multi-level emergent behaviours in designed component-based systems and agent-based simulations". *From System*

Complexity to Emergent Properties. Ed. by Aziz-Alaoui, Moulay & Bertelle, Cyrille. Springer Berlin Heidelberg. Chap. 4, pp. 101–114 (cited on p. 38).

Chen, Eric Olav, Alexis Ghersengorin & Sami Petersen (2024a). *Imperfect Recall and AI Delegation*. Tech. rep. 30. Global Priorities Institute Working Paper Series (cited on p. 36).

Chen, Hao, Bhiksha Raj, Xing Xie & Jindong Wang (2024b). "On Catastrophic Inheritance of Large Foundation Models". *ArXiv* abs/2402.01909 (cited on p. 26).

Chen, Junzhe, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He & Lijie Wen (2024c). "LLMArena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments". *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Ku, Lun-Wei, Martins, Andre & Srikumar, Vivek, pp. 13055–13077 (cited on p. 45).

Chen, Weize, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun & Jie Zhou (2024d). "AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors". *The Twelfth International Conference on Learning Representations* (cited on p. 38).

Cheung, Yun Kuen & Georgios Piliouras (2020). "Chaos, Extremism and Optimism: Volume Analysis of Learning in Games". *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Larochelle, Hugo, Ranzato, Marc'Aurelio, Hadsell, Raia, Balcan, Maria-Florina & Lin, Hsuan-Tien (cited on pp. 31, 32).

Chopra, Samir & Laurence F. White (2011). *A Legal Theory for Autonomous Artificial Agents*. University of Michigan Press (cited on pp. 17, 36, 46).

Chotibut, Thiparat, Fryderyk Falniowski, Michał Misiurewicz & Georgios Piliouras (2020). "The route to chaos in routing games: When is price of anarchy too optimistic?" *Advances in Neural Information Processing Systems* 33, pp. 766–777 (cited on p. 31).

Christiano, Paul (2018). *Clarifying "AI alignment"*. AI Alignment. URL: https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6 (cited on p. 43).

Christiano, Paul, Buck Shlegeris & Dario Amodei (2018). "Supervising Strong Learners by Amplifying Weak Experts". *arXiv:1810.08575* (cited on pp. 18, 20, 41, 43).

Christiano, Paul, Mark Xu & Ajeya Cotra (2021). *Eliciting Latent Knowledge*. Alignment Forum. URL: https://www.alignmentforum.org/posts/qHCDysDnvhteW7kRd/arc-s-first-technical-report-eliciting-latent-knowledge (cited on pp. 18, 20, 41, 43).

Christoffersen, Phillip J. K., Andreas A. Haupt & Dylan Hadfield-Menell (2023). "Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL". *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 448–456 (cited on p. 35).

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi & Michele Starnini (2021). "The echo chamber effect on social media". *Proceedings of the National Academy of Sciences* 118.9, e2023301118 (cited on p. 47).

Cisco (2023). *What is a distributed denial-of-service (DDoS) attack?* URL: https://www.cloudflare.com/learning/ddos/what-is-a-ddos-attack/ (cited on p. 39).

Clark, Herbert H. & Susan E. Brennan (1991). "Grounding in communication". *Perspectives on socially shared cognition*. American Psychological Association, pp. 127–149 (cited on p. 12).

Clifton, Jesse (2020). *Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda*. Tech. rep. Center on Long-Term Risk (cited on p. 8).

Clune, Jeff (2019). "AI-GAs: AI-generating Algorithms, an Alternate Paradigm for Producing General Artificial Intelligence". *arXiv:1905.10985* (cited on p. 27).

Cohen, Alon, Argyrios Deligkas & Moran Koren (2023). "Learning approximately optimal contracts". *Theoretical Computer Science* 980, p. 114219 (cited on p. 35).

Cohen, Reuven & Shlomo Havlin (2010). *Complex networks: structure, robustness and function.* Cambridge university press (cited on p. 23).

Colognese, Paul & Arun Jose (2023). *High-level interpretability: detecting an AI's objectives.* Alignment Forum. URL: https://www.alignmentforum.org/posts/tFYGdq9ivjA3rdaS2/high-level-interpretability-detecting-an-ai-s-objectives (cited on pp. 19, 38).

Commission, U.S. Commodity Futures Trading & U.S. Securities & Exchange Commission (2010). *Findings Regarding the Market Events of May 6, 2010.* Tech. rep. (cited on p. 31).

Conitzer, Vincent, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde & William S. Zwicker (2024). "Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback". *Forty-first International Conference on Machine Learning* (cited on p. 47).

Conitzer, Vincent & Caspar Oesterheld (2023). "Foundations of Cooperative AI". *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence.* Vol. 37. 13, pp. 15359–15367 (cited on pp. 8, 16, 22, 30, 36, 37, 43).

Conitzer, Vincent & Tuomas Sandholm (2004). "Computational criticisms of the revelation principle". *Proceedings of the 5th ACM conference on Electronic commerce*, pp. 262–263 (cited on p. 23).

Cooper, Emery, Caspar Oesterheld & Vincent Conitzer (2025). "Characterising Simulation-Based Program Equilibria". *Proceedings of the Thirty-Ninth Annual AAAI Conference on Artificial Intelligence* (cited on p. 16).

Cooper, Russell W., Douglas V. DeJong, Robert Forsythe & Thomas W. Ross (1990). "Selection Criteria in Coordination Games: Some Experimental Results". *The American Economic Review* 80.1, pp. 218–233 (cited on p. 11).

Copeland, Dale C. (2000). *The Origins of Major War.* Cornell University Press (cited on p. 34).

Couzin, Iain (2007). "Collective minds". *Nature* 445.7129, pp. 715–715 (cited on p. 36).

Cover, Thomas M. & Joy A. Thomas (2005). *Elements of Information Theory.* Wiley (cited on p. 17).

Crawford, John David (1991). "Introduction to bifurcation theory". *Reviews of Modern Physics* 63.4, pp. 991–1037 (cited on p. 32).

Crawford, Vincent P. & Joel Sobel (1982). "Strategic Information Transmission". *Econometrica* 50.6, p. 1431 (cited on p. 12).

Critch, Andrew, Michael Dennis & Stuart Russell (2022). "Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory". *arXiv:2208.07006* (cited on pp. 16, 36).

Critch, Andrew & David Krueger (2020). "AI Research Considerations for Human Existential Safety (ARCHES)". *arXiv:2006.04948* (cited on pp. 5, 9, 43).

Critch, Andrew & Stuart Russell (2023). "TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI". *arXiv:2306.06924* (cited on pp. 9, 44).

Crootof, Rebecca, Margot E. Kaminski & W. Nicholson I. I. Price (2023). "Humans in the Loop". *Vanderbilt Law Review* 76, p. 429 (cited on p. 35).

Cross, Logan, Violet Xiang, Agam Bhatia, Daniel L. K. Yamins & Nick Haber (2025). "Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models". *The Thirteenth International Conference on Learning Representations* (cited on p. 13).

Csernatoni, Raluca (2024). *Can Democracy Survive the Disruptive Power of AI?* Tech. rep. Carnegie Endowment for International Peace (cited on pp. 34, 47).

CSIS (2023). *Significant Cyber Incidents — Strategic Technologies Program.* Center for Strategic & Internation Studies. URL: https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents (cited on p. 39).

Curtis, Samuel, Ravi Iyer, Cameron Domenico Kirk-Giannini, Victoria Krakovna, David Krueger, Nathan Lambert, Bruno Marnette, Colleen McKenzie, Julian Michael, Evan Miyazono, Noyuri Mima, Aviv

Ovadya, Luke Thorburn & Deger Turan (2024). *Research Agenda for Sociotechnical Approaches to AI Safety*. Tech. rep. AI Objectives Institute (cited on pp. 5, 43, 44).

Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel (2021). "Cooperative AI: machines must learn to find common ground". *Nature* 593.7857, pp. 33–36 (cited on pp. 8, 30).

Dafoe, Allan, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson & Thore Graepel (2020). "Open Problems in Cooperative AI". *arXiv:2012.08630* (cited on pp. 5, 8, 13, 43).

Dahia, Simranjeet Singh & Claudia Szabo (2024). "Detecting Emergent Behavior in Complex Systems: A Machine Learning Approach". *Proceedings of the 38th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pp. 81–87 (cited on p. 38).

Dai, Jessica, Paula Gradu, Inioluwa Deborah Raji & Benjamin Recht (2025). "From Individual Experience to Collective Evidence: A Reporting-Based Framework for Identifying Systemic Harms". *arXiv:2502.08166* (cited on pp. 26, 45).

Daley, Brendan & Philipp Sadowski (2017). "Magical thinking: A representation result". *Theoretical Economics* 12.2, pp. 909–956 (cited on p. 17).

Dalrymple, David "davidad", Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing & Joshua Tenenbaum (2024). "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems". *arXiv:2405.06624* (cited on pp. 20, 43).

Darley, John M. & Bibb Latané (1968). "Bystander intervention in emergencies: diffusion of responsibility." *Journal of personality and social psychology* 8.4p1, p. 377 (cited on p. 48).

Daskalakis, Constantinos, Alan Deckelbaum & Anthony Kim (2011). "Near-Optimal No-Regret Algorithms for Zero-Sum Games". *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 235–254 (cited on p. 15).

Daskalakis, Constantinos, Dylan J Foster & Noah Golowich (2020). "Independent Policy Gradient Methods for Competitive Reinforcement Learning". *Advances in Neural Information Processing Systems*. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F. & Lin, H. Vol. 33, pp. 5527–5540 (cited on p. 15).

Dawes, Robyn & David Messick (2000). "Social Dilemmas". *International Journal of Psychology* 35, pp. 111– (cited on p. 13).

De Clippel, Geoffroy, Herve Moulin & Nicolaus Tideman (2008). "Impartial division of a dollar". *Journal of Economic Theory* 139.1, pp. 176–191 (cited on p. 48).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia

Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang & Zhen Zhang (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". *arXiv:2501.12948* (cited on p. 26).

Dennett, Daniel (1971). "Intentional Systems". *Journal of Philosophy* 68.4, pp. 87–106 (cited on p. 37).

Dennis, Michael, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch & Sergey Levine (2020). "Emergent complexity and zero-shot transfer via unsupervised environment design". *Advances in neural information processing systems* 33, pp. 13049–13061 (cited on p. 29).

Desai, Nishant, Andrew Critch & Stuart J. Russell (2018). "Negotiable reinforcement learning for pareto optimal sequential decision-making". *Adv. Neural Inf. Process. Syst.* 31 (cited on p. 47).

Devillers, Laurence, Françoise Fogelman-Soulié & Ricardo Baeza-Yates (2021). "AI & Human Values: Inequalities, Biases, Fairness, Nudge, and Feedback Loops". *Reflections on Artificial Intelligence for Humanity*, pp. 76–89 (cited on p. 48).

DiGiovanni, Anthony & Jesse Clifton (2023). "Commitment games with conditional information disclosure". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 5, pp. 5616–5623 (cited on pp. 22, 42).

DiGiovanni, Anthony, Jesse Clifton & Nicolas Macé (2024). "Safe Pareto Improvements for Expected Utility Maximizers in Program Games". *arXiv:2403.05103* (cited on p. 22).

DiGiovanni, Anthony, Nicolas Macé & Jesse Clifton (2022). "Evolutionary Stability of Other-Regarding Preferences Under Complexity Costs" (cited on p. 28).

Disarmament Affairs, United Nations Office for (2023). "Lethal Autonomous Weapon Systems (LAWS)". *United Nations Office for Disarmament Affairs* (cited on p. 17).

Domingos, Elias Fernández, Francisco C. Santos & Tom Lenaerts (2023). "EGTtools: Evolutionary game dynamics in Python". *Iscience* 26.4 (cited on p. 30).

Domingos, Elias Fernández, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C. Burguillo, Francisco C. Santos & Tom Lenaerts (2022). "Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma". *Scientific Reports* 12.1 (cited on p. 46).

Doumbouya, Moussa Koulako Bala, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky & Christopher D. Manning (2024). *h4rm3l: A Dynamic Benchmark of Composable Jailbreak Attacks for LLM Safety Assessment* (cited on p. 25).

Dowlin, Nathan, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig & John Wernsing (2016). "CryptoNets: applying neural networks to encrypted data with high throughput and accuracy". *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pp. 201–210 (cited on p. 36).

Draguns, Andis, Andrew Gritsevskiy, Sumeet Ramesh Motwani & Christian Schroeder de Witt (2024). "Unelicitable Backdoors via Cryptographic Transformer Circuits". *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (cited on p. 41).

Drechsler, Itamar, Alexi Savov, Philipp Schnabl & Olivier Wang (2023). *Deposit Franchise Runs*. Tech. rep. 31138. National Bureau of Economic Research (cited on p. 14).

Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. Tech. rep. 2019-1. Future of Humanity Institute, University of Oxford. Chap. 20 (cited on pp. 7, 20, 37, 43, 44).

Drexler, K. Eric (2022). *Applying superintelligence without collusion*. Alignment Forum. URL: https://www.alignmentforum.org/posts/HByDKLLdaWEcA2QQD/applying-superintelligence-without-collusion (cited on pp. 5, 18, 43).

Driessen, Theo (1988). *Cooperative games, solutions and applications*. Kluwer Academic Publishers (cited on p. 16).

Du, Yali, Joel Z. Leibo, Usman Islam, Richard Willis & Peter Sunehag (2023). "A Review of Cooperation in Multi-Agent Learning". *arXiv:2312.05162* (cited on p. 12).

Duan, Jinhao, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen & Kaidi Xu (2024). "GTBench: Uncovering the Strategic Reasoning Capabilities of LLMs via Game-Theoretic Evaluations". *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (cited on p. 16).

Dughmi, Shaddin (2019). "On the hardness of designing public signals". *Games and Economic Behavior* 118, pp. 609–625 (cited on p. 22).

Duong, Manh Hong & The Anh Han (2021). "Cost efficiency of institutional incentives for promoting cooperation in finite populations". *Proceedings of the Royal Society A* 477.2254, p. 20210568 (cited on p. 30).

Dwork, Cynthia (2006). "Differential privacy". *International colloquium on automata, languages, and programming*. Springer, pp. 1–12 (cited on p. 48).

Elliott, Matthew, Benjamin Golub & Matthew O. Jackson (2014). "Financial Networks and Contagion". *American Economic Review* 104.10, pp. 3115–3153 (cited on p. 23).

Ellsberg, Daniel (1968). *The theory and practice of blackmail*. Rand Corporation Santa Monica, CA (cited on p. 14).

Ely, Jeffrey C. & Balazs Szentes (2023). *Natural Selection of Artificial Intelligence*. Tech. rep. National University of Singapore (cited on p. 28).

Emmons, Scott, Caspar Oesterheld, Andrew Critch, Vincent Conitzer & Stuart Russell (2022). "For Learning in Symmetric Teams, Local Optima are Global Nash Equilibria". *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Chaudhuri, Kamalika, Jegelka, Stefanie, Song, Le, Szepesvari, Csaba, Niu, Gang & Sabato, Sivan. Vol. 162, pp. 5924–5943 (cited on pp. 12, 16).

Epoch (2023). *ML trends*. Epoch. URL: https://epochai.org/trends (cited on p. 25).

EU (2024). "REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024: laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)". *Official Journal of the European Union* L series.2024/1689. Text with EEA relevance (cited on p. 45).

Evans, Owain, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti & William Saunders (2021). "Truthful AI: Developing and Governing AI That Does Not Lie". *arXiv:2110.06674* (cited on pp. 18, 21, 23).

Everitt, Tom, Ryan Carey, Eric D. Langlois, Pedro A. Ortega & Shane Legg (2021). "Agent Incentives: A Causal Perspective". *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11487–11495 (cited on pp. 37, 38).

Ezrachi, Ariel & Maurice E. Stucke (2017). "Artificial Intelligence & Collusion: When Computers Inhibit Competition". *University of Illinois Law Review* (cited on p. 18).

Farrell, Joseph & Matthew Rabin (1996a). "Cheap Talk". *The Journal of Economic Perspectives* 10.3, pp. 103–118 (cited on p. 12).

— (1996b). "Cheap talk". *Journal of Economic perspectives* 10.3, pp. 103–118 (cited on p. 12).

Fearon, James D. (1995). “Rationalist explanations for war”. *Int. Organ.* 49.3, pp. 379–414 (cited on pp. 16, 21, 34, 43).

Feng, Xiachong, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma & Lingpeng Kong (2024). “A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios”. *arXiv:2412.03920* (cited on p. 30).

Fernandes, Pedro M., Francisco C. Santos & Manuel Lopes (2020). “Adoption Dynamics and Societal Impact of AI Systems in Complex Networks”. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 258–264 (cited on p. 26).

Fernando, Chrisantha, Dylan Banarse, Henryk Michalewski, Simon Osindero & Tim Rocktäschel (2023). “Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution”. *arXiv:2309.16797* (cited on p. 27).

Ferrara, Emilio (2023). “Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models”. *arXiv:2304.03738* (cited on p. 28).

Ferreira, Fernando G. D. C., Amir H. Gandomi & Rodrigo T. N. Cardoso (2021). “Artificial Intelligence Applied to Stock Market Trading: A Review”. *IEEE Access* 9, pp. 30898–30917 (cited on pp. 4, 5, 23).

Fickinger, Arnaud, Simon Zhuang, Dylan Hadfield-Menell & Stuart Russell (2020). “Multi-principal assistance games”. *arXiv:2007.09540* (cited on p. 47).

Figueiredo, João, Afonso Carvalho, Daniel Castro, Daniel Gonçalves & Nuno Santos (2024). “On the Feasibility of Fully AI-automated Vishing Attacks”. *arXiv:2409.13793* (cited on p. 40).

Fish, Sara, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira & Manuel Wüthrich (2023). “Generative Social Choice”. *arXiv:2309.01291* (cited on p. 46).

Fish, Sara, Yannai A. Gonczarowski & Ran I. Shorrer (2024). “Algorithmic Collusion by Large Language Models”. *arXiv:2404.00806* (cited on p. 19).

Fishkin, James, Nikhil Garg, Lodewijk Gelauff, Ashish Goel, Sukolsak Sakshuwong, Alice Siu, Kamesh Munagala & Sravya Yandamuri (2019). “Deliberative Democracy with the Online Deliberation Platform”. *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing* (cited on p. 46).

Foerster, Jakob, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel & Igor Mordatch (2018). “Learning with Opponent-learning Awareness”. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130 (cited on p. 16).

Foerster, Jakob N., Yannis M. Assael, Nando de Freitas & Shimon Whiteson (2016). “Learning to Communicate with Deep Multi-agent Reinforcement Learning”. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2145–2153 (cited on pp. 18, 21).

Fontana, Magda & Pietro Terna (2015). “From Agent-Based Models to Network Analysis (and Return): The Policy-Making Perspective”. *Journal on Policy and Complex Systems* (cited on p. 26).

Forsberg, Erika (2008). “Polarization and ethnic conflict in a widened strategic setting”. *Journal of Peace Research* 45.2, pp. 283–300 (cited on p. 23).

Fort, Stanislav (2023). *GPT-4(V) vs Dall-e 3*. Twitter. URL: https://twitter.com/stanislavfort/status/1713603557046276334 (cited on p. 33).

Fosong, Elliot, Arrasy Rahman, Ignacio Carlucho & Stefano V. Albrecht (2022). “Few-Shot Teamwork”. *arXiv:2207.09300* (cited on p. 22).

Foxabbott, Jack, Sam Deverett, Kaspar Senft, Samuel Dower & Lewis Hammond (2023). “Defining and Mitigating Collusion in Multi-Agent Systems”. *Multi-Agent Security Workshop at NeurIPS* (cited on p. 19).

Franzmeyer, Tim, Stephen Marcus McAleer, Joao F. Henriques, Jakob Nicolaus Foerster, Philip Torr, Adel Bibi & Christian Schroeder de Witt (2023). “Illusory Attacks: Information-theoretic detectability matters in adversarial attacks”. *The Twelfth International Conference on Learning Representations* (cited on pp. 41, 42).

Friedenberg, Meir & Joseph Y. Halpern (2019). "Blameworthiness in multi-agent settings". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 525–532 (cited on pp. 7, 48).

— (2023). "Joint Behavior and Common Belief". *Electronic Proceedings in Theoretical Computer Science* 379, pp. 221–232 (cited on p. 48).

Friston, Karl J, Maxwell J D Ramstead, Alex B Kiefer, Alexander Tschantz, Christopher L Buckley, Mahault Albarracin, Riddhi J Pitliya, Conor Heins, Brennan Klein, Beren Millidge, Dalton A R Sakthivadivel, Toby St Clere Smithe, Magnus Koudahl, Safae Essafi Tremblay, Capm Petersen, Kaiser Fung, Jason G Fox, Steven Swanson, Dan Mapes & Gabriel René (2022). "Designing Ecosystems of Intelligence from First Principles". *Collective Intelligence, 3(1), 2024* 3.1 (cited on p. 38).

Fu, Yao, Hao Peng, Tushar Khot & Mirella Lapata (2023). "Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback". *arXiv:2305.10142* (cited on p. 30).

Gabriel, Iason, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac & James Manyika (2024). "The Ethics of Advanced AI Assistants". *arXiv:2404.16244* (cited on pp. 7, 8, 28, 34, 42, 46, 47).

Gaertner, Wulf (2010). *A primer in social choice theory*. Rev. ed., reprinted. Literaturverz. S. [201] - 207. - Literaturangaben. Oxford Univ. Press. 218 pp. (cited on p. 16).

Galla, Tobias & J. Doyne Farmer (2013). "Complex Dynamics in Learning Complicated Games". *Proceedings of the National Academy of Sciences* 110.4, pp. 1232–1236 (cited on p. 32).

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan & Jack Clark (2022). "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned". *arXiv:2208.07858* (cited on p. 29).

Gao, Jianxi, Baruch Barzel & Albert-László Barabási (2016). "Universal resilience patterns in complex networks". *Nature* 530.7590, pp. 307–312 (cited on pp. 23, 25).

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser & Connor Leahy (2020). "The Pile: An 800GB Dataset of Diverse Text for Language Modeling". *ArXiv* abs/2101.00027 (cited on p. 26).

Gavrilets, Sergey (2015). "Collective action problem in heterogeneous groups". *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1683, p. 20150016 (cited on pp. 43, 44).

Ge, Yingqiang, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou & Yongfeng Zhang (2020). "Understanding echo chambers in e-commerce recommender systems". *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 2261–2270 (cited on p. 47).

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii & Kate Crawford (2021). "Datasheets for datasets". *Communications of the ACM* 64.12, pp. 86–92 (cited on p. 45).

Gemp, Ian M., Kevin R. McKee, Richard Everett, Edgar A. Duéñez-Guzmán, Yoram Bachrach, David Balduzzi & Andrea Tacchetti (2022). "D3C: Reducing the Price of Anarchy in Multi-Agent Learning". *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022,* *Auckland, New Zealand, May 9-13, 2022*. Ed. by Faliszewski, Piotr, Mascardi, Viviana, Pelachaud, Catherine & Taylor, Matthew E., pp. 498–506 (cited on p. 16).

Gerstgrasser, Matthias & David Parkes (2023). "Oracles & followers: stackelberg equilibria in deep multi-agent reinforcement learning". *Proceedings of the 40th International Conference on Machine Learning* (cited on pp. 15, 33).

Gerstgrasser, Matthias, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho & Sanmi Koyejo (2024). "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data". *First Conference on Language Modeling* (cited on p. 32).

Ghodsi, Zahra, Tianyu Gu & Siddharth Garg (2017). "SafetyNets: Verifiable Execution of Deep Neural Networks on an Untrusted Cloud". *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4675–4684 (cited on p. 36).

Gibbard, Allan (1973). "Manipulation of Voting Schemes: A General Result". *Econometrica* 41.4, p. 587 (cited on p. 23).

Gilbert, Nigel (2019). *Agent-based models*. Sage Publications (cited on p. 30).

Gilbert, Thomas Krendl, Nathan Lambert, Sarah Dean, Tom Zick & Aaron Snoswell (2022). "Reward Reports for Reinforcement Learning". *arXiv:2204.10817* (cited on p. 33).

Gilbert, Thomas Krendl, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell & Soham Mehta (2023). "Reward Reports for Reinforcement Learning". *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 84–130 (cited on p. 45).

Gillies, Donald B. (1959). "Solutions to general non-zero-sum games". *Contributions to the Theory of Games* 4.40, pp. 47–85 (cited on p. 16).

Gleave, Adam, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine & Stuart Russell (2020). "Adversarial Policies: Attacking Deep Reinforcement Learning". *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* (cited on pp. 15, 29, 44).

Gleick, James (1998). *Chaos. Making a new science*. Originally published: New York: Viking, 1987; London: Heinemann, 1988. Vintage (cited on p. 31).

Goddard, Kate, Abdul Roudsari & Jeremy C. Wyatt (2012). "Automation bias: a systematic review of frequency, effect mediators, and mitigators". *Journal of the American Medical Informatics Association* 19.1, pp. 121–127 (cited on p. 35).

Goel, Shashwat, Joschka Struber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge & Jonas Geiping (2025). "Great Models Think Alike and this Undermines AI Oversight". *arXiv:2502.04313* (cited on pp. 18, 24, 26, 43).

Goldstein, Josh A., Renee DiResta, Girish Sastry, Micah Musser, Matthew Gentzel & Katerina Sedova (2023). *Generative Language Models and AutomatedInfluence Operations: Emerging Threats and Potential Mitigations*. Tech. rep. Georgetown University's Center for Security, Emerging Technology, OpenAI, and Stanford Internet Observatory (cited on p. 21).

Google DeepMind (2024). *Introducing Gemini 2.0: our new AI model for the agentic era*. URL: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message (cited on pp. 4, 25, 37).

Gordon, Deborah M. (1996). "The organization of work in social insect colonies". *Nature* 380.6570, pp. 121–124 (cited on p. 36).

Gordon, Mitchell L., Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto & Michael S. Bernstein (2022). "Jury Learning: Integrating Dissenting Voices into Machine Learning Models". *CHI Conference on Human Factors in Computing Systems*, pp. 1–19 (cited on p. 47).

Gosmar, Diego, Deborah A. Dahl, Emmett Coin & David Attwater (2024). *AI Multi-Agent Interoperability Extension for Managing Multiparty Conversations* (cited on p. 42).

Grafen, Alan (1990). “Biological signals as handicaps”. *Journal of theoretical biology* 144.4, pp. 517–546 (cited on pp. 28, 30).

Green, Ben (2022). “The flaws of policies requiring human oversight of government algorithms”. *Computer Law & Security Review* 45, p. 105681 (cited on p. 35).

Green, Ben & Yiling Chen (2020). “Algorithm-in-the-Loop Decision Making”. *Proceedings of the AAAI Conference on Artificial Intelligence* 34.09, pp. 13663–13664 (cited on p. 35).

Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan & Fabien Roger (2023). “AI Control: Improving Safety Despite Intentional Subversion”. *arXiv:2312.06942* (cited on pp. 7, 18, 20, 23, 36, 41, 43).

Griffin, Charlie, Louis Thomson, Buck Shlegeris & Alessandro Abate (2024). “Games for AI Control: Models of Safety Evaluations of AI Deployment Protocols”. *arXiv:2409.07985* (cited on p. 36).

Grossman, Sanford J. (1981). “The Informational Role of Warranties and Private Disclosure about Product Quality”. *The Journal of Law and Economics* 24.3, pp. 461–483 (cited on p. 22).

Grupen, Niko, Natasha Jaques, Been Kim & Shayegan Omidshafiei (2022). “Concept-based Understanding of Emergent Multi-Agent Behavior”. *Deep Reinforcement Learning Workshop at NeurIPS* (cited on p. 38).

Gu, Xiangming, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang & Min Lin (2024). “Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast”. *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Salakhutdinov, Ruslan, Kolter, Zico, Heller, Katherine, Weller, Adrian, Oliver, Nuria, Scarlett, Jonathan & Berkenkamp, Felix. Vol. 235, pp. 16647–16672 (cited on pp. 24, 25, 41, 42).

Guo, Hao, Chen Shen, Shuyue Hu, Junliang Xing, Pin Tao, Yuanchun Shi & Zhen Wang (2023). “Facilitating cooperation in human-agent hybrid populations through autonomous agents”. *Iscience* 26.11 (cited on p. 30).

Hadfield-Menell, Dylan, McKane Andrus & Gillian Hadfield (2019). “Legible normativity for AI alignment: The value of silly rules”. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 115–121 (cited on p. 12).

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel & Stuart Russell (2016). “Cooperative Inverse Reinforcement Learning”. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3916–3924 (cited on p. 43).

Haghtalab, Nika, Nicole Immorlica, Brendan Lucier, Markus Mobius & Divyarthi Mohan (2024). “Communicating with Anecdotes”. *15th Innovations in Theoretical Computer Science Conference, ITCS*. Ed. by Guruswami, Venkatesan. Vol. 287, 57:1–57:2 (cited on pp. 21, 22).

Haghtalab, Nika, Michael Jordan & Eric Zhao (2022). “On-demand sampling: Learning optimally from multiple distributions”. *Advances in Neural Information Processing Systems* 35, pp. 406–419 (cited on p. 48).

Haghtalab, Nika, Mingda Qiao & Kunhe Yang (2025). “Platforms for Efficient and Incentive-Aware Collaboration”. *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, pp. 2607–2628 (cited on p. 22).

Halawi, Danny, Alexander Wei, Eric Wallace, Tony Tong Wang, Nika Haghtalab & Jacob Steinhardt (2024). “Covert Malicious Finetuning: Challenges in Safeguarding LLM Adaptation”. *Forty-first International Conference on Machine Learning, ICML 2024* (cited on pp. 41, 42).

Halpern, Joseph Y. & Max Kleiman-Weiner (2018). “Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility”. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by McIlraith, Sheila A. & Weinberger, Kilian Q., pp. 1853–1860 (cited on pp. 37, 38).

Halpern, Joseph Y. & Rafael Pass (2018). “Game theory with translucent players”. *International Journal of Game Theory* 47.3, pp. 949–976 (cited on pp. 17, 22, 36).

Hammond, Lewis & Sam Adam-Day (2025). "Neural Interactive Proofs". *The Thirteenth International Conference on Learning Representations*. Forthcoming (cited on pp. 42, 45).

Hammond, Lewis, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate & Michael Wooldridge (2023). "Reasoning about causality in games". *Artificial Intelligence* 320, p. 103919 (cited on p. 38).

Han, The Anh (2022). "Emergent behaviours in multi-agent systems with Evolutionary Game Theory." *AI Communications* 35.4 (cited on p. 28).

Han, The Anh, Manh Hong Duong & Matjaz Perc (2024). "Evolutionary mechanisms that promote cooperation may not promote social welfare". *Journal of the Royal Society Interface* 21.220, p. 20240547 (cited on p. 30).

Han, The Anh, Cedric Perret & Simon T. Powers (2021). "When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games". *Cognitive Systems Research* 68, pp. 111–124 (cited on pp. 30, 36).

Hao, Jianye, Jun Sun, Guangyong Chen, Zan Wang, Chao Yu & Zhong Ming (2017). "Efficient and Robust Emergence of Norms through Heuristic Collective Learning". *ACM Trans. Auton. Adapt. Syst.* 12.4 (cited on p. 12).

Hardin, G. (1968). "The tragedy of the commons". *Science* 162, pp. 1243–1248 (cited on pp. 13, 14).

Hardt, Moritz, Eric Price & Nati Srebro (2016). "Equality of opportunity in supervised learning". *Advances in neural information processing systems* 29 (cited on p. 48).

Hardy, Amelia, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M. Asmar, Sanmi Koyejo, Michael S. Bernstein & Mykel J. Kochenderfer (2024). "More than Marketing? On the Information Value of AI Benchmarks for Practitioners". *arXiv:2412.05520* (cited on p. 45).

Harrenstein, Paul, Felix Brandt & Felix Fischer (2007). "Commitment and extortion". *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 1–8 (cited on p. 14).

Harrington, Joseph E. (2019). "Developing Competition Law for Collusion by Autonomous Artificial Agents". *Journal of Competition Law & Economics* 14.3, pp. 331–363 (cited on pp. 17, 18).

Harsanyi, John C. & Reinhard Selten (1988). "A general theory of equilibrium selection in games". *MIT Press Books* 1 (cited on p. 16).

Hart, Sergiu & Andreu Mas-Colell (2000). "A Simple Adaptive Procedure Leading to Correlated Equilibrium". *Econometrica* 68.5, pp. 1127–1150 (cited on p. 17).

Hasan, Syed Mhamudul, Alaa M. Alotaibi, Sajedul Talukder & Abdur R. Shahid (2024). "Distributed Threat Intelligence at the Edge Devices: A Large Language Model-Driven Approach". *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1496–1497 (cited on p. 42).

Hawkins, Robert X. D., Noah D. Goodman & Robert L. Goldstone (2019). "The emergence of social norms and conventions". *Trends in cognitive sciences* 23.2, pp. 158–169 (cited on p. 30).

Hazell, Julian (2023). "Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns". arXiv.org. *arXiv:2305.06972* (cited on p. 40).

He, Yifeng, Ethan Wang, Yuyang Rong, Zifei Cheng & Hao Chen (2024). "Security of AI Agents". *arXiv:2406.08689* (cited on p. 41).

Hendrycks, Dan (2023). "Natural Selection Favors AIs over Humans". *arXiv:2303.16200* (cited on pp. 9, 28).

Hendrycks, Dan, Nicholas Carlini, John Schulman & Jacob Steinhardt (2021). "Unsolved Problems in ML Safety". *arXiv:2109.13916* (cited on pp. 4, 5, 7, 8, 43).

Hendrycks, Dan, Mantas Mazeika & Thomas Woodside (2023). "An Overview of Catastrophic AI Risks". *arXiv:2306.12001* (cited on p. 9).

Ho, Chien-Ju, Aleksandrs Slivkins & Jennifer Wortman Vaughan (2014). "Adaptive contract design for crowdsourcing markets: bandit algorithms for repeated principal-agent problems". *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 359–376 (cited on p. 35).

Hofbauer, Josef & Karl Sigmund (1998). *Evolutionary games and population dynamics*. Cambridge university press (cited on p. 30).

Hoffman, David E. (2009). *The Dead Hand. The untold story of the Cold War arms race and its dangerous legacy*. Doubleday (cited on p. 35).

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals & Laurent Sifre (2022). *Training Compute-Optimal Large Language Models*. arXiv:2203.15556 [cs] (cited on p. 25).

Hofstadter, Douglas (1983). "Dilemmas for Superrational Thinkers, Leading Up to a Luring Lottery". *Scientific America* 248.6 (cited on p. 17).

Holling, C. S. (1973). "Resilience and Stability of Ecological Systems". *Annual Review of Ecology and Systematics* 4, pp. 1–23 (cited on p. 31).

Horowitz, M. C. (2021). "When speed kills: Lethal autonomous weapon systems, deterrence and stability". *Emerging Technologies and International Stability* (cited on p. 14).

Hossain, Safwan, Evi Micha & Nisarg Shah (2021). "Fair algorithms for multi-agent multi-armed bandits". *Advances in Neural Information Processing Systems* 34, pp. 24005–24017 (cited on p. 48).

Hossain, Safwan, Andjela Mladenovic & Nisarg Shah (2020). "Designing fairly fair classifiers via economic fairness notions". *Proceedings of The Web Conference 2020*, pp. 1559–1569 (cited on p. 48).

Hossain, Safwan, Tonghan Wang, Tao Lin, Yiling Chen, David C. Parkes & Haifeng Xu (2024). "Multi-Sender Persuasion: A Computational Perspective". *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Salakhutdinov, Ruslan, Kolter, Zico, Heller, Katherine, Weller, Adrian, Oliver, Nuria, Scarlett, Jonathan & Berkenkamp, Felix. Vol. 235, pp. 18944–18971 (cited on p. 22).

Howard, J. V. (1988). "Cooperation in the Prisoner's Dilemma". *Theory and Decision* 24.3, pp. 203–213 (cited on pp. 16, 22, 34, 36).

Hu, Donghui, Liang Wang, Wenjie Jiang, Shuli Zheng & Bin Li (2018). "A Novel Image Steganography Method via Deep Convolutional Generative Adversarial Networks". *IEEE Access* 6, pp. 38303–38314 (cited on p. 18).

Hu, Hengyuan, Adam Lerer, Alex Peysakhovich & Jakob Foerster (2020). ""Other-Play " for zero-shot coordination". *Proceedings of the 37th International Conference on Machine Learning* (cited on pp. 12, 16, 22).

Hu, Shengran, Cong Lu & Jeff Clune (2024). "Automated Design of Agentic Systems". *arXiv:2408.08435* (cited on p. 27).

Huang, Jen-tse, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Maarten Sap & Michael R. Lyu (2024). "On the Resilience of Multi-Agent Systems with Malicious Agents". *arXiv:2408.00989* (cited on pp. 26, 29, 41, 42, 44).

Huang, Ling, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein & J. D. Tygar (2011). "Adversarial machine learning". *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (cited on pp. 18, 20, 43).

Huang, Minyi, Roland P. Malhamé & Peter E. Caines (2006). "Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle". *Communications in Information & Systems* 6.3, pp. 221–252 (cited on p. 33).

Huang, Saffron & Divya Siddarth (2023). "Generative AI and the Digital Commons". *arXiv:2303.11074* (cited on pp. 24, 46).

Hubinger, Evan (2020). *Clarifying inner alignment terminology*. Alignment Forum. URL: https://www.alignmentforum.org/posts/SzecSPYxqRa5GCaSF/clarifying-inner-alignment-terminology (cited on p. 43).

Hughes, Edward, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul & Tim Rocktaschel (2024). "Open-Endedness is Essential for Artificial Superhuman Intelligence". *arXiv:2406.04268* (cited on p. 29).

Hughes, Edward, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff & Thore Graepel (2018). "Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas". *Advances in Neural Information Processing Systems 31*. Ed. by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. Vol. 31, pp. 3326–3336 (cited on p. 16).

Hutter, Frank, Lars Kotthoff & Joaquin Vanschoren, eds. (2019). *Automated Machine Learning*. Springer International Publishing (cited on p. 37).

Irving, Geoffrey, Paul Christiano & Dario Amodei (2018). "AI Safety Via Debate". *arXiv:1805.00899* (cited on pp. 7, 18, 20, 23, 41, 43).

Islam, Mohammad Saiful, Mehmet Kuzu & Murat Kantarcioglu (2012). "Access Pattern disclosure on Searchable Encryption: Ramification, Attack and Mitigation". *19th Annual Network and Distributed System Security Symposium, NDSS 2012, San Diego, California, USA, February 5-8, 2012* (cited on p. 40).

Itzhak, Itay, Gabriel Stanovsky, Nir Rosenfeld & Yonatan Belinkov (2023). "Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias". *arXiv:2308.00225* (cited on p. 28).

Ivanov, Dmitry, Ilya Zisman & Kirill Chernyshev (2023). "Mediated Multi-Agent Reinforcement Learning". *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 49–57 (cited on p. 22).

Jackson, Joshua Conrad, Virginia K. Choi & Michele J. Gelfand (2019). "Revenge: A Multilevel Review and Synthesis". *Annu. Rev. Psychol.* 70, pp. 319–345 (cited on p. 28).

Jackson, Matthew O. & Yves Zenou (2015). "Chapter 3 - Games on Networks". *Handbook of Game Theory with Economic Applications*. Ed. by Young, H. Peyton & Zamir, Shmuel. Vol. 4. Elsevier, pp. 95–163 (cited on p. 23).

Jaderberg, Max, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio G. Castañeda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu & Thore Graepel (2019). "Human-level performance in 3D multiplayer games with population-based reinforcement learning". *Science* 364.6443, pp. 859–865 (cited on p. 27).

Jagadeesan, Meena, Michael Jordan, Jacob Steinhardt & Nika Haghtalab (2023a). "Improved Bayes Risk Can Yield Reduced Social Welfare Under Competition". *Thirty-seventh Conference on Neural Information Processing Systems* (cited on pp. 5, 43).

Jagadeesan, Meena, Michael I. Jordan & Nika Haghtalab (2023b). "Competition, Alignment, and Equilibria in Digital Marketplaces". *Proceedings of the AAAI Conference on Artificial Intelligence* 37.5, pp. 5689–5696 (cited on p. 18).

Jaques, Natasha, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo & Nando de Freitas (2019). "Social Influence As Intrinsic Motivation for Multi-agent Deep Reinforcement Learning". *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Chaudhuri, Kamalika & Salakhutdinov, Ruslan. Vol. 97, pp. 3040–3049 (cited on p. 16).

Jarrett, Daniel, Miruna Pislar, Michael Tessler, Michiel Bakker, Raphael Koster, Jan Balaguer, Romuald Elie, Christopher Summerfield & Andrea Tacchetti (2023). "Language Agents as Digital Representatives in Collective Decision-Making". *NeurIPS Workshop on Foundation Models for Decision Making* (cited on p. 46).

Jennings, Nick R. (1993). "Specification and Implementation of a Belief-Desire-Joint-Intention Architecture for Collaborative Problem Solving". *International Journal of Cooperative Information Systems* 02.03, pp. 289–318 (cited on p. 48).

Jervis, Robert (2017). *Perception and Misperception in International Politics: New Edition.* Princeton University Press (cited on p. 14).

Jia, Hengrui, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran & Nicolas Papernot (2021). "Proof-of-Learning: Definitions and Practice". *2021 IEEE Symposium on Security and Privacy (SP)* (cited on p. 36).

Jiang, Minqi, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette & Tim Rocktäschel (2021). "Replay-guided adversarial environment design". *Advances in Neural Information Processing Systems* 34, pp. 1884–1897 (cited on p. 29).

Jiang, Ray, Silvia Chiappa, Tor Lattimore, András György & Pushmeet Kohli (2019). "Degenerate feedback loops in recommender systems". *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390 (cited on p. 47).

Johnson, Bonnie (2021a). "Artificial intelligence systems: unique challenges for defense applications". *Acquisition Research Program* (cited on p. 14).

Johnson, Dominic D. P. (2004). *Overconfidence and War: The Havoc and Glory of Positive Illusions.* Harvard University Press (cited on p. 14).

Johnson, James (2020). "Artificial intelligence: a threat to strategic stability". *Strategic Studies Quarterly* 14.1, pp. 16–39 (cited on p. 14).

— (2021b). "Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age". *European Journal of International Security* 7.3, pp. 337–359 (cited on p. 33).

Jones, Erik, Anca Dragan & Jacob Steinhardt (2024). "Adversaries Can Misuse Combinations of Safe Models". *arXiv:2406.14595* (cited on pp. 5, 18, 20, 40, 42, 44).

Jones, Erik & Jacob Steinhardt (2022). "Capturing failures of large language models via human cognitive biases". Ed. by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. & Oh, A., pp. 11785–11799 (cited on p. 28).

Ju, Tianjie, Yiting Wang, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang & Gongshen Liu (2024). "Flooding Spread of Manipulated Knowledge in LLM-Based Multi-Agent Communities". *arXiv:2407.07791* (cited on pp. 24, 25, 41, 42).

Justesen, Niels, Ruben Rodriguez Torrado, Philip Bontrager, Ahmed Khalifa, Julian Togelius & Sebastian Risi (2018). "Illuminating generalization in deep reinforcement learning through procedural level generation". *arXiv:1806.10729* (cited on p. 29).

Kairouz, Peter, Sewoong Oh & Pramod Viswanath (2015). "The composition theorem for differential privacy". *International conference on machine learning.* PMLR, pp. 1376–1385 (cited on p. 48).

Kalai, Ehud & Meir Smorodinsky (1975). "Other solutions to Nash's bargaining problem". *Econometrica: Journal of the Econometric Society*, pp. 513–518 (cited on p. 16).

Kamenica, Emir & Matthew Gentzkow (2011). "Bayesian Persuasion". *American Economic Review* 101.6, pp. 2590–2615 (cited on p. 22).

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei (2020). *Scaling Laws for Neural Language Models.* arXiv:2001.08361 [cs, stat] (cited on p. 25).

Kapoor, Sayash, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir & Arvind Narayanan (2024). "AI Agents That Matter". *arXiv:2407.01502* (cited on pp. 8, 45).

Kasirzadeh, Atoosa (2024a). "Plurality of value pluralism and AI value alignment". *Pluralistic Alignment Workshop at NeurIPS* (cited on p. 47).

— (2024b). "Two Types of AI Existential Risk: Decisive and Accumulative". *arXiv:2401.07836* (cited on pp. 9, 44).

Kaunert, Christian & Ethem Ilbiz (2021). “Cyber-attacks: what is hybrid warfare and why is it such a threat?” *The Conversation* (cited on p. 39).

Kay, Jackie, Atoosa Kasirzadeh & Shakir Mohamed (2024). “Epistemic Injustice in Generative AI”. *arXiv:2408.11441* (cited on pp. 21, 24, 47).

Kenton, Zachary, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott & Tom Everitt (2022). “Discovering Agents”. *arXiv:2208.08345* (cited on pp. 37, 38).

Khan, Akbir, Timon Willi, Newton Kwan, Andrea Tacchetti, Chris Lu, Edward Grefenstette, Tim Rocktäschel & Jakob Foerster (2023). “Scaling Opponent Shaping to High Dimensional Games”. *arXiv:2312.12568* (cited on p. 16).

Khlaaf, Heidy (2023). *Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems.* Trail of Bits. URL: https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf (cited on pp. 39, 42).

Kinniment, Megan, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes & Paul Christiano (2023). *Evaluating Language-Model Agents on Realistic Autonomous Tasks.* ARC Evals. URL: https://evals.alignment.org/language-model-pilot-report (cited on pp. 29, 45).

Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum & Tom Goldstein (2023). “On the Reliability of Watermarks for Large Language Models”. *arXiv:2306.04634* (cited on p. 20).

Kirilenko, Andrei A., Albert S. Kyle, Mehrdad Samadi & Tugkan Tuzun (2017). “The Flash Crash: High-Frequency Trading in an Electronic Market”. *The Journal of Finance* 72.3, pp. 967–998 (cited on pp. 5, 31, 33).

Kirk, Hannah, Bertie Vidgen, Paul Rottger & Scott Hale (2023). “The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models”. *Socially Responsible Language Modelling Research* (cited on p. 43).

Klein, Timo (2021). “Autonomous algorithmic collusion: Q-learning under sequential pricing”. *The RAND Journal of Economics* 52.3, pp. 538–558 (cited on pp. 18, 19).

Kleinberg, Jon & Manish Raghavan (2021). “Algorithmic monoculture and social welfare”. *Proceedings of the National Academy of Sciences* 118.22. Publisher: Proceedings of the National Academy of Sciences, e2018340118 (cited on p. 45).

Koessler, Leonie & Jonas Schuett (2023). *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries.* arXiv:2307.08823 [cs] (cited on p. 45).

Kokotajlo, Daniel (2019). *The Commitment Races problem.* Alignment Forum. URL: https://www.alignmentforum.org/posts/brXr7PJ2W4Na2EW2q/the-commitment-races-problem (cited on p. 35).

Kollock, Peter (1998). “Social dilemmas: The anatomy of cooperation”. *Annual review of sociology* 24.1, pp. 183–214 (cited on p. 13).

Kolt, Noam (2024). “Governing AI Agents”. *SSRN Electronic Journal* (cited on pp. 17, 35, 36, 46).

Kolt, Noam, Michal Shur-Ofry & Reuven Cohen (2025). “Lessons from complexity theory for AI governance”. *arXiv:2502.00012* (cited on pp. 17, 44).

Konda, Vijay R. & John N. Tsitsiklis (2000). “Actor-critic Algorithms”. *Advances in Neural Information Processing Systems 12.* Ed. by Solla, S. A., Leen, T. K. & Müller, K., pp. 1008–1014 (cited on p. 37).

Kong, Yuqing & Grant Schoenebeck (2019). “An Information Theoretic Framework For Designing Information Elicitation Mechanisms That Reward Truth-telling”. *ACM Transactions on Economics and Computation* 7.1, pp. 1–33 (cited on p. 22).

Konrad, Kai A. & Florian Morath (2012). “Evolutionarily stable in-group favoritism and out-group spite in intergroup conflict”. *J. Theor. Biol.* 306, pp. 61–67 (cited on p. 28).

Köster, Raphael, Kevin R. McKee, Richard Everett, Laura Weidinger, William S. Isaac, Edward Hughes, Edgar A. Duéñez-Guzmán, Thore Graepel, Matthew Botvinick & Joel Z. Leibo (2020). “Model-free

Conventions in Multi-agent Reinforcement Learning with Heterogeneous Preferences". *arXiv:2010.09054* (cited on pp. 12, 15).

Kovařík, Vojtěch, Caspar Oesterheld & Vincent Conitzer (2023). "Game Theory with Simulation of Other Players". *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (cited on p. 36).

— (2024). "Recursive Joint Simulation in Games". *arXiv:2402.08128* (cited on p. 36).

Kreps, Sarah & Doug Kriner (2023). "How AI Threatens Democracy". *Journal of Democracy* 34.4, pp. 122–31 (cited on pp. 34, 47).

Krier, Sébastien (2023). *An agents economy.* URL: https://www.aipolicyperspectives.com/p/an-agents-economy (cited on p. 34).

Kubík, Aleš (2003). "Toward a Formalization of Emergence". *Artificial Life* 9.1, pp. 41–65 (cited on p. 38).

Kulveit, Jan (2023). *Personal Communication* (cited on p. 25).

Kulveit, Jan, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger & David Duvenaud (2025). "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development". *arXiv:2501.16946* (cited on p. 44).

Kunegis, Jérôme, Marcel Blattner & Christine Moser (2013). "Preferential attachment in online networks: measurement and explanations". *Web Science Conference* (cited on p. 25).

Kushner, Harold J. & G. George Yin (2003). *Stochastic Modelling and Applied Probability.* Springer-Verlag (cited on p. 33).

Laird, Burgess (2020). "The Risks of Autonomous Weapons Systems for Crisis Stability and Conflict Escalation in Future US-Russia Confrontations". *Rand Corperation: Objective Analysis. Effective Solutions* (cited on p. 14).

Lamparth, Max, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider & Harold Trinkunas (2024). "Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations". *arXiv:2403.03407* (cited on p. 15).

Lamport, Leslie, Robert Shostak & Marshall Pease (1982). "The Byzantine Generals Problem". *ACM Transactions on Programming Languages and Systems* 4.3, pp. 382–401 (cited on p. 41).

Lasry, Jean-Michel & Pierre-Louis Lions (2007). "Mean field games". *Japanese Journal of Mathematics* 2.1, pp. 229–260 (cited on p. 33).

Lauffer, Niklas, Ameesh Shah, Micah Carroll, Michael D. Dennis & Stuart Russell (2023). "Who Needs to Know? Minimal Knowledge for Optimal Coordination". *Proceedings of the 40th International Conference on Machine Learning.* ISSN: 2640-3498, pp. 18599–18613 (cited on p. 21).

Laurito, Walter, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm & Jan Kulveit (2024). "AI AI Bias: Large Language Models Favor Their Own Generated Content" (cited on pp. 24, 26).

Lazar, Seth & Alondra Nelson (2023). "AI safety on whose terms?" *Science* 381.6654. Publisher: American Association for the Advancement of Science, pp. 138–138 (cited on pp. 5, 43, 44).

Lazaridou, Angeliki & Marco Baroni (2020). "Emergent Multi-Agent Communication in the Deep Learning Era". *arXiv:2006.02419* (cited on pp. 18, 21, 37).

Lee, Donghyun & Mo Tiwari (2024). "Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems". *arXiv:2410.07283* (cited on pp. 5, 24, 25, 41, 42).

Legg, Shane & Marcus Hutter (2007). "Universal intelligence: A definition of machine intelligence". *Minds and Machines* 17.4, pp. 391–444 (cited on p. 37).

Lehman, Joel, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh & Kenneth O. Stanley (2022). "Evolution through Large Models". *arXiv:2206.08896* (cited on p. 27).

Leibo, Joel Z., Edgar A. Dueñez-Guzman, Alexander Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch & Thore Graepel (2021). "Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot". *Proceedings of the 38th Inter-*

national Conference on Machine Learning. Ed. by Meila, Marina & Zhang, Tong. Vol. 139, pp. 6187–6199 (cited on pp. 12, 23, 29, 32).

Leibo, Joel Z., Edward Hughes, Marc Lanctot & Thore Graepel (2019). "Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research". *arXiv:1903.00742* (cited on pp. 27, 28).

Leibo, Joel Z., Julien Perolat, Edward Hughes, Steven Wheelwright, Adam H. Marblestone, Edgar Duéñez-Guzmán, Peter Sunehag, Iain Dunning & Thore Graepel (2018). "Malthusian Reinforcement Learning". *arXiv:1812.07019* (cited on p. 27).

Leibo, Joel Z., Alexander Sasha Vezhnevets, Manfred Diaz, John P. Agapiou, William A. Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A. Duéñez-Guzmán, William S. Isaac, Georgios Piliouras, Stanley M. Bileschi, Iyad Rahwan & Simon Osindero (2024). "A theory of appropriateness with applications to generative artificial intelligence". *arXiv:2412.19010* (cited on p. 12).

Leibo, Joel Z., Vinicius Zambaldi, Marc Lanctot, Janusz Marecki & Thore Graepel (2017). "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473 (cited on pp. 14, 16).

Leike, Jan, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini & Shane Legg (2018). "Scalable Agent Alignment Via Reward Modeling: A Research Direction". *arXiv:1811.07871* (cited on pp. 18, 20, 41, 43).

Leisten, Matthew (2021). "Algorithmic competition, with humans". Working Paper (cited on p. 19).

Leonardos, Stefanos & Georgios Piliouras (2022). "Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory". *Artificial Intelligence* 304, p. 103653 (cited on pp. 32, 33).

Leonardos, Stefanos, Iosif Sakos, Costas Courcoubetis & Georgios Piliouras (2020). "Catastrophe by Design in Population Games: Destabilizing WastefulLocked-In Technologies". *Web and Internet Economics - 16th International Conference, WINE 2020, Beijing, China, December 7-11, 2020, Proceedings*. Ed. by Chen, Xujin, Gravin, Nikolai, Hoefer, Martin & Mehta, Ruta. Vol. 12495, p. 473 (cited on p. 32).

Lerer, Adam & Alexander Peysakhovich (2019). "Learning Existing Social Conventions via Observationally Augmented Self-Play". *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Ed. by Conitzer, Vincent, Hadfield, Gillian K. & Vallor, Shannon, pp. 107–114 (cited on pp. 12, 23).

Lermen, Simon & Ondřej Kvapil (2023). "Exploring the Robustness of Model-Graded Evaluations and Automated Interpretability". *arXiv:2312.03721* (cited on p. 51).

Letchford, Joshua, Dmytro Korzhyk & Vincent Conitzer (2013). "On the value of commitment". *Autonomous Agents and Multi-Agent Systems* 28.6, pp. 986–1016 (cited on pp. 34, 47).

Levermore, Patrick (2023). *AI Safety Bounties*. Rethink Priorities. URL: https://rethinkpriorities.org/research-area/ai-safety-bounties/ (cited on p. 44).

Levinstein, B. A. & Daniel A. Herrmann (2023). "Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks". *Philosophical Studies* (cited on pp. 22, 23).

Lewis, David (1979). "Prisoners' Dilemma is a Newcomb Problem". *Philosophy & Public Affairs* 8.3, pp. 235–240 (cited on p. 17).

Lewontin, Richard C. (1958). "A general method for investigating the equilibrium of gene frequency in a population". *Genetics* 43.3, p. 419 (cited on p. 30).

Li, Huao, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis & Katia Sycara (2023a). "Theory of Mind for Multi-Agent Collaboration via Large Language Models". *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Bouamor, Houda, Pino, Juan & Bali, Kalika, pp. 180–192 (cited on p. 13).

Li, Wenhao, Dan Qiao, Baoxiang Wang, Xiangfeng Wang, Wei, Hao Shen, Bo Jin & Hongyuan Zha (2025). "Multi-Agent Credit Assignment with Pretrained Language Models". *The 28th International Conference on Artificial Intelligence and Statistics* (cited on p. 12).

Li, Yang, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang & Wei Pan (2023b). "Cooperative open-ended learning framework for zero-shot coordination". *Proceedings of the 40th International Conference on Machine Learning* (cited on p. 23).

Liang, Paul Pu, Chiyu Wu, Louis-Philippe Morency & Ruslan Salakhutdinov (2021). "Towards Understanding and Mitigating Social Biases in Language Models". *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Meila, Marina & Zhang, Tong. Vol. 139, pp. 6565–6576 (cited on p. 28).

Liao, Hung-Jen, Chun-Hung Richard Lin, Ying-Chih Lin & Kuang-Yuan Tung (2013). "Intrusion detection system: A comprehensive review". *Journal of Network and Computer Applications* 36.1, pp. 16–24 (cited on p. 41).

Lima, Dafni (2017). "Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law". *South Carolina Law Review* 69.3, pp. 677–696 (cited on pp. 17, 36, 46).

Lin, Tao & Ce Li (2024). "Information Design with Unknown Prior". *arXiv:2410.05533* (cited on p. 22).

Lior, Anat (2019). "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy". *Mitchell Hamline Law Review* 46.5, pp. 1043–1102 (cited on pp. 17, 36, 46).

Liu, Yiqi, Nafise Moosavi & Chenghua Lin (2024). "LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores". *Findings of the Association for Computational Linguistics ACL 2024*, pp. 12688–12701 (cited on pp. 25, 26).

Löwenheim, Oded & Gadi Heimann (2008). "Revenge in International Politics". *Security Studies* 17.4, pp. 685–724 (cited on p. 28).

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune & David Ha (2024a). "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery". *arXiv:2408.06292* (cited on p. 37).

Lu, Christopher, Timon Willi, Christian A. Schröder de Witt & Jakob N. Foerster (2022). "Model-Free Opponent Shaping". *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Chaudhuri, Kamalika, Jegelka, Stefanie, Song, Le, Szepesvári, Csaba, Niu, Gang & Sabato, Sivan. Vol. 162, pp. 14398–14411 (cited on p. 16).

Lu, Yikang, Alberto Aleta, Chunpeng Du, Lei Shi & Yamir Moreno (2024b). "LLMs and generative agent-based models for complex systems research". *Physics of Life Reviews* (cited on p. 30).

Luo, Yi, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu & Chen Lin (2024). "From Intention To Implementation: Automating Biomedical Research via LLMs". *arXiv:2412.09429* (cited on pp. 37, 44).

Lupu, Andrei, Brandon Cui, Hengyuan Hu & Jakob Foerster (2021). "Trajectory Diversity for Zero-Shot Coordination". *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Meila, Marina & Zhang, Tong. Vol. 139, pp. 7204–7213 (cited on p. 23).

Lupu, Andrei & Doina Precup (2020). "Gifting in Multi-Agent Reinforcement Learning". *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 789–797 (cited on p. 15).

Lyon, Pamela (2011). "To Be or Not To Be: Where Is Self-Preservation in Evolutionary Theory?" *The Major Transitions in Evolution Revisited*. The MIT Press. Chap. 6, pp. 105–126 (cited on p. 38).

Lyu, Xin (2022). "Composition theorems for interactive differential privacy". *Advances in Neural Information Processing Systems* 35, pp. 9700–9712 (cited on p. 48).

Maas, Matthijs M. (2018). "Regulating for "Normal AI Accidents": Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment". *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 223–228 (cited on pp. 5, 26, 31, 44, 46).

MacDermott, Matt, James Fox, Francesco Belardinelli & Tom Everitt (2024). "Measuring Goal-Directedness". *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (cited on pp. 37, 38).

Mahowald, Kyle, Anna Ivanova, Idan Blank, Nancy Kanwisher, Joshua Tenenbaum & Evelina Fedorenko (2023). *Dissociating language and thought in large language models: a cognitive perspective* (cited on p. 12).

Malone, Thomas W. & Michael S. Bernstein (2022). *Handbook of collective intelligence*. MIT press (cited on p. 37).

Manheim, David (2019). "Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence". *Big Data and Cognitive Computing* 3.2, p. 21 (cited on pp. 5, 8, 43).

Mankowitz, Daniel J., Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, Thomas Köppe, Kevin Millikin, Stephen Gaffney, Sophie Elster, Jackson Broshear, Chris Gamble, Kieran Milan, Robert Tung, Minjae Hwang, Taylan Cemgil, Mohammadamin Barekatain, Yujia Li, Amol Mandhane, Thomas Hubert, Julian Schrittwieser, Demis Hassabis, Pushmeet Kohli, Martin Riedmiller, Oriol Vinyals & David Silver (2023). "Faster sorting algorithms discovered using deep reinforcement learning". *Nature* 618.7964, pp. 257–263 (cited on p. 37).

Manson, Katrina (2023). "The US Military Is Taking Generative AI Out for a Spin" (cited on p. 14).

— (2024). "AI Warfare Is Already Here". *Bloomberg* (cited on pp. 4, 5, 14).

Maoz, Zeev (2012). "Preferential Attachment, Homophily, and the Structure of International Networks, 1816–2003". *Conflict Management and Peace Science* 29, pp. 341–369 (cited on p. 25).

Marks, Luke, Amir Abdullah, Luna Mendez, Rauno Arike, Philip Torr & Fazl Barez (2023). "Interpreting Reward Models in RLHF-Tuned Language Models Using Sparse Autoencoders". *arXiv:2310.08164* (cited on pp. 19, 38).

Marro, Samuele, Emanuele La Malfa, Jesse Wright, Guohao Li, Nigel Shadbolt, Michael Wooldridge & Philip Torr (2024). "A Scalable Communication Protocol for Networks of Large Language Models". *arXiv:2410.11905* (cited on pp. 12, 42, 45).

Martínez, Gonzalo, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juárez & Rik Sarkar (2023). "Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet". *arXiv:2306.06130* (cited on pp. 26, 32).

Martins, Paulo, Leonel Sousa & Artur Mariano (2017). "A Survey on Fully Homomorphic Encryption: An Engineering Perspective". *ACM Computing Surveys* 50.6, pp. 1–33 (cited on p. 36).

Mashayekhi, Mehdi, Nirav Ajmeri, George F. List & Munindar P. Singh (2022). "Prosocial Norm Emergence in Multi-agent Systems". *ACM Trans. Auton. Adapt. Syst.* 17.1–2 (cited on p. 12).

Mathew, Yohan, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope & Nandi Schoots (2024). "Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs". *arXiv:2410.03768* (cited on pp. 18, 19).

Mayer, Dominik, Johannes Feldmaier & Hao Shen (2016). "Reinforcement Learning in Conflicting Environments for Autonomous Vehicles". *International Workshop on Robotics in the 21st Century: Challenges and Promises* (cited on pp. 17, 23).

Mayorkas, Alejandro N. (2024). *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure*. Report. PDF available online. U.S. Department of Homeland Security (cited on pp. 4, 23).

Mazeika, Mantas, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang & Dan Hendrycks (2025). "Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs". *arXix:2502.08640* (cited on p. 28).

McAfee, R. Preston (1984). "Effective Computability in Economic Decisions" (cited on pp. 16, 22, 34, 36).

McElreath, Richard, Robert Boyd & PeterJ Richerson (2003). "Shared norms and the evolution of ethnic markers". *Current anthropology* 44.1, pp. 122–130 (cited on p. 12).

McKee, Kevin R., Andrea Tacchetti, Michiel A. Bakker, Jan Balaguer, Lucy Campbell-Gillingham, Richard Everett & Matthew Botvinick (2023). “Scaffolding cooperation in human groups with deep reinforcement learning”. *Nature Human Behaviour* 7, pp. 1787–1796 (cited on p. 13).

McNally, Luke & Andrew L. Jackson (2013). “Cooperation Creates Selection for Tactical Deception”. *Proceedings of the Royal Society B: Biological Sciences* 280.1762, p. 20130699 (cited on p. 28).

Mehra, Salil K. (2016). “Antitrust and the Robo-Seller: Competition in the Time of Algorithms”. *Minnesota Law Review* 204 (cited on p. 18).

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan (2021). “A survey on bias and fairness in machine learning”. *ACM computing surveys (CSUR)* 54.6, pp. 1–35 (cited on p. 47).

Meta (2025). *Building Toward a Smarter, More Personalized Assistant.* URL: https://about.fb.com/news/2025/01/building-toward-a-smarter-more-personalized-assistant/ (cited on pp. 25, 37).

Meulemans, Alexander, Seijin Kobayashi, Johannes von Oswald, Nino Scherrer, Eric Elmoznino, Blake Richards, Guillaume Lajoie, Blaise Agüera y Arcas & João Sacramento (2024). “Multi-agent cooperation through learning-aware policy gradients”. *arXiv:2410.18636* (cited on p. 16).

Meyersohn, Nathaniel (2024). “How Red Lobster’s misguided endless shrimp promotion drove it into bankruptcy”. *CNN* (cited on p. 34).

Mialon, Grégoire, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun & Thomas Scialom (2023). “Augmented Language Models: a Survey”. *arXiv:2302.07842* (cited on p. 24).

Micha, Evi & Nisarg Shah (2020). “Proportionally fair clustering revisited”. *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020).* Schloss Dagstuhl-Leibniz-Zentrum für Informatik (cited on p. 48).

Michaud, Eric J., Adam Gleave & Stuart Russell (2020). “Understanding Learned Reward Functions”. *Deep RL Workshop at NeurIPS* (cited on pp. 19, 38).

Microsoft (2024). *Introducing Azure AI Agent Service.* URL: https://techcommunity.microsoft.com/blog/azure-ai-services-blog/introducing-azure-ai-agent-service/4298357 (cited on pp. 25, 37).

Milgrom, Paul R. (1981). “Good News and Bad News: Representation Theorems and Applications”. *The Bell Journal of Economics* 12.2, pp. 380–391 (cited on p. 22).

Miller, Nolan, Paul Resnick & Richard Zeckhauser (2005). “Eliciting Informative Feedback: The Peer-Prediction Method”. *Management Science* 51.9, pp. 1359–1373 (cited on p. 22).

Mini, Ulisse, Peli Grietzer, Mrinank Sharma, Austin Meek, Monte MacDiarmid & Alexander Matt Turner (2023). “Understanding and Controlling a Maze-Solving Policy Network”. *arXiv:2310.08043* (cited on pp. 19, 38).

Mirza, M. Usman, Andries Richter, Egbert H. van Nes & Marten Scheffer (2019). “Technology driven inequality leads to poverty and resource depletion”. *Ecological Economics* 160, pp. 215–226 (cited on p. 47).

Mitchell, Eric, Charles Lin, Antoine Bosselut, Chelsea Finn & Christopher D. Manning (2022). “Fast Model Editing at Scale”. *International Conference on Learning Representations* (cited on p. 20).

Mitchell, Margaret, Avijit Ghosh, Alexandra Sasha Luccioni & Giada Pistilli (2025). “Fully Autonomous AI Agents Should Not be Developed”. *arXiv:2502.02649* (cited on pp. 8, 43, 45).

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji & Timnit Gebru (2019). “Model Cards for Model Reporting”. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (cited on pp. 33, 45).

Mogul, Jeffrey C. (2006). “Emergent (mis)behavior vs. complex software systems”. *ACM SIGOPS Operating Systems Review* 40.4, pp. 293–304 (cited on pp. 5, 8, 31, 36).

Morris-Martin, Andreasa, Marina De Vos & Julian Padget (2019). “Norm emergence in multiagent systems: a viewpoint paper”. *Autonomous Agents and Multi-Agent Systems* 33.6, pp. 706–749 (cited on p. 12).

Motter, Adilson E. & Ying-Cheng Lai (2002). “Cascade-based attacks on complex networks”. *Physical Review E* 66.6. arXiv:cond-mat/0301086, p. 065102 (cited on p. 41).

Motwani, Sumeet, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond & Christian Schroeder de Witt (2024). “Secret Collusion among AI Agents: Multi-Agent Deception via Steganography”. *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (cited on pp. 5, 19, 41).

Myerson, Roger B. & Mark A. Satterthwaite (1983). “Efficient mechanisms for bilateral trading”. *Journal of economic theory* 29.2, pp. 265–281 (cited on pp. 20, 21).

Nadeem, Moin, Anna Bethke & Siva Reddy (2020). “StereoSet: Measuring stereotypical bias in pretrained language models”. *arXiv:2004.09456* (cited on p. 28).

Nagarajan, Sai Ganesh, David Balduzzi & Georgios Piliouras (2020). “From Chaos to Order: Symmetry and Conservation Laws in Game Dynamics”. *Proceedings of the 37th International Conference on Machine Learning*, pp. 7186–7196 (cited on pp. 32, 33).

Nangia, Nikita, Clara Vania, Rasika Bhalerao & Samuel R. Bowman (2020). “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models”. *arXiv:2010.00133* (cited on p. 28).

Narang, Adhyyan, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel & Lillian J. Ratliff (2023). “Multiplayer Performative Prediction: Learning in Decision-Dependent Games”. *Journal of Machine Learning Research* 24.202, pp. 1–56 (cited on p. 32).

Nash, John (1950). “The bargaining problem”. *Econometrica: Journal of the econometric society*, pp. 155–162 (cited on p. 16).

— (1951). “Non-Cooperative Games”. *Annals of Mathematics* 54.2. Publisher: Annals of Mathematics, pp. 286–295 (cited on p. 11).

National Artificial Intelligence Research Resource Task Force (2023). *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem*. Tech. rep. National Artificial Intelligence Initiative (cited on p. 45).

NCSC (2023). *The threat from commercial cyber proliferation*. National Cyber Security Centre. URL: https://www.ncsc.gov.uk/report/commercial-cyber-proliferation-assessment (visited on 09/28/2023) (cited on p. 39).

— (n.d.). *SBOMs and the importance of inventory*. National Cyber Security Centre. URL: https://www.ncsc.gov.uk/blog-post/sboms-and-the-importance-of-inventory (cited on p. 42).

Neslon, Alondra (2023). *Thick Alignment*. Ethics in AI Annual Lecture, University of Oxford (cited on p. 43).

Newman, M. E. J. (2003). “The Structure and Function of Complex Networks”. *SIAM Review* 45.2, pp. 167–256 (cited on p. 25).

Newman, Mark & Mark Newman (2018). *Networks*. Second Edition, Second Edition. Oxford University Press (cited on p. 23).

Newman, Richard B. (2024). “FTC Announces Final Rule Imposing Civil Penalties for Fake Consumer Reviews and Testimonials”. *The National Law review* (cited on p. 40).

Ngo, Richard, Lawrence Chan & Sören Mindermann (2022). “The alignment problem from a deep learning perspective”. *arXiv:2209.00626* (cited on pp. 7, 29).

Nguyen, Dung, Hung Le, Kien Do, Sunil Gupta, Svetha Venkatesh & Truyen Tran (2024a). “Diversifying Training Pool Predictability for Zero-shot Coordination: A Theory of Mind Approach”. *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pp. 166–174 (cited on p. 22).

Nguyen, Minh-Duong, Quang Vinh Do, Zhaohui Yang, Won-Joo Hwang & Quoc-Viet Pham (2024b). "Distortion Resilience for Goal-Oriented Semantic Communication". *IEEE Transactions on Mobile Computing*, pp. 1–12 (cited on p. 24).

Normann, Hans-Theo & Martin Sternberg (2023). "Human-algorithm interaction: Algorithmic pricing in hybrid laboratory markets". *European Economic Review* 152, p. 104347 (cited on p. 19).

Nowak, Martin A. (2006). "Five rules for the evolution of cooperation". *Science* 314.5805, pp. 1560–1563 (cited on pp. 28, 30).

NSF (2023). *Safe Learning-Enabled Systems*. National Science Foundation. URL: https://www.nsf.gov/funding/opportunities/safe-learning-enabled-systems (cited on p. 44).

Nyborg, Karine, John M. Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W. Neil Adger, Kenneth J. Arrow, Scott Barrett, Stephen Carpenter, F. Stuart Chapin, Anne-Sophie Crépin, Gretchen Daily, Paul Ehrlich, Carl Folke, Wander Jager, Nils Kautsky, Simon A. Levin, Ole Jacob Madsen, Stephen Polasky, Marten Scheffer, Brian Walker, Elke U. Weber, James Wilen, Anastasios Xepapadeas & Aart de Zeeuw (2016). "Social norms as solutions". *Science* 354.6308, pp. 42–43 (cited on p. 12).

O'toole, Eamonn, Vivek Nallur & Siobhán Clarke (2017). "Decentralised Detection of Emergence in Complex Adaptive Systems". *ACM Transactions on Autonomous and Adaptive Systems* 12.1, pp. 1–31 (cited on p. 38).

Oesterheld, Caspar (2016). "Formalizing preference utilitarianism in physical world models". *Synthese* 193.9, pp. 2747–2759 (cited on p. 37).

— (2018). "Robust Program Equilibrium". *Theory and Decision* 86.1, pp. 143–159 (cited on pp. 16, 22, 36).

— (2021). "Approval-directed agency and the decision theory of Newcomb-like problems". *Synthese* 198.Suppl 27, pp. 6491–6504 (cited on p. 17).

Oesterheld, Caspar & Vincent Conitzer (2022). "Safe Pareto improvements for delegated game playing". *Autonomous Agents and Multi-Agent Systems* 36.2, p. 46 (cited on pp. 16, 22, 46).

Oesterheld, Caspar, Emery Cooper, Miles Kodama, Linh Chi Nguyen & Ethan Perez (2024a). "A dataset of questions on decision-theoretic reasoning in Newcomb-like problems". *arXiv:2411.10588* (cited on p. 17).

Oesterheld, Caspar, Johannes Treutlein, Emery Cooper & Rubi Hudson (2023). "Incentivizing honest performative predictions with proper scoring rules". *Uncertainty in Artificial Intelligence*. PMLR, pp. 1564–1574 (cited on p. 21).

Oesterheld, Caspar, Johannes Treutlein, Roger B. Grosse, Vincent Conitzer & Jakob Foerster (2024b). "Similarity-based cooperative equilibrium". *Advances in Neural Information Processing Systems* 36 (cited on pp. 13, 17, 23).

Okasha, Samir (2006). *Evolution and the levels of selection*. Clarendon Press (cited on p. 27).

— (2018). *Agents and Goals in Evolution*. Oxford University Press (cited on p. 38).

Oldenburg, Ninell & Tan Zhi-Xuan (2024). "Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games". *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1510–1520 (cited on p. 15).

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish & Chris Olah (2022). "In-context Learning and Induction Heads". *arXiv:2209.11895* (cited on p. 32).

Omidshafiei, Shayegan, Jason Pazis, Christopher Amato, Jonathan P. How & John Vian (2017). "Deep Decentralized Multi-task Multi-agent Reinforcement Learning under Partial Observability". *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 2681–2690 (cited on pp. 10, 12, 15).

Omohundro, Stephen M. (2008). “The Basic AI Drives”. *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pp. 483–492 (cited on pp. 29, 38).

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard & Wojciech Marian Czarnecki (2021). “Open-Ended Learning Leads to Generally Capable Agents” (cited on p. 27).

OpenAI (2023a). *GPT-4 System Card*. URL: https://cdn.openai.com/papers/gpt-4-system-card.pdf (cited on p. 18).

— (2023b). *Schelling Point Eval*. GitHub. URL: https://github.com/openai/evals/tree/main/evals/elsuite/schelling_point (cited on p. 13).

— (2023c). *Text Compression Eval*. GitHub. URL: https://github.com/openai/evals/blob/main/evals/elsuite/text_compression (cited on pp. 18, 19).

— (2024). *OpenAI o1 System Card*. OpenAI. URL: https://openai.com/index/openai-o1-system-card/ (cited on p. 47).

— (2025). *Introducing Operator*. URL: https://openai.com/index/introducing-operator/ (cited on pp. 4, 25, 37).

Option Alpha (2025). *Financial History* (cited on p. 31).

Ord, Toby (2020). *The precipice: existential risk and the future of humanity*. Hachette Books (cited on p. 9).

Oroojlooy, Afshin & Davood Hajinezhad (2022). “A review of cooperative multi-agent deep reinforcement learning”. *Applied Intelligence* 53.11, pp. 13677–13722 (cited on pp. 10, 12, 15).

Orseau, Laurent, Simon McGregor McGill & Shane Legg (2018). “Agents and Devices: A Relative Definition of Agency”. *arXiv:1805.12387* (cited on p. 37).

Ostrom, Elinor (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, p. 280 (cited on pp. 13, 41, 43, 44).

Ovadya, Aviv (2023). *'Generative CI' through Collective Response Systems*. arXiv:2302.00672 [cs] (cited on p. 46).

Pacchiardi, Lorenzo, Alex James Chan, Sören Mindermann, Ilan Moscovitz, Alexa Yue Pan, Yarin Gal, Owain Evans & Jan M. Brauner (2024). “How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions”. *The Twelfth International Conference on Learning Representations* (cited on p. 23).

Padalkar, Abhishek, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J. Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R. Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh-hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu & Zichen Jeff Cui (2023). "Open X-Embodiment: Robotic Learning Datasets and RT-X Models". *arXiv:2310.08864* (cited on pp. 11, 51).

Paes Leme, Renato, Georgios Piliouras, Jon Schneider, Kelly Spendlove & Song Zuo (2024). "Complex Dynamics in Autobidding Systems". *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 75–100 (cited on p. 32).

Palaiopanos, Gerasimos, Ioannis Panageas & Georgios Piliouras (2017). "Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos". *Advances in Neural Information Processing Systems* 30 (cited on p. 32).

Palantir (2025). *AIP for Defense* (cited on pp. 4, 5, 14, 15).

Pan, Alexander, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons & Dan Hendrycks (2023). "Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark". *International Conference on Machine Learning*, pp. 26837–26867 (cited on p. 29).

Panickssery, Arjun, Samuel R. Bowman & Shi Feng (2024). "LLM Evaluators Recognize and Favor Their Own Generations". *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (cited on pp. 25, 26).

Papadimitriou, Christos & Georgios Piliouras (2019). "Game dynamics as the meaning of a game". *ACM SIGecom Exchanges* 16.2, pp. 53–63 (cited on p. 30).

Papoudakis, Georgios, Filippos Christianos, Arrasy Rahman & Stefano V. Albrecht (2019). "Dealing with Non-Stationarity in Multi-Agent Deep Reinforcement Learning". *arXiv:1906.04737* (cited on p. 32).

Pardoe, David, Peter Stone, Maytal Saar-Tsechansky & Kerem Tomak (2006). "Adaptive mechanism design". *Proceedings of the 8th international conference on Electronic commerce The new e-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet - ICEC '06* (cited on pp. 15, 33).

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang & Michael S. Bernstein (2023a). "Generative agents: Interactive simulacra of human behavior". *arXiv:2304.03442* (cited on p. 38).

Park, Peter S., Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks (2023b). "AI Deception: A Survey of Examples, Risks, and Potential Solutions". *arXiv:2308.14752* (cited on p. 18).

Park, Peter S., Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks (2024). "AI deception: A survey of examples, risks, and potential solutions". *Patterns* 5.5, p. 100988 (cited on pp. 21, 23).

Parker-Holder, Jack, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette & Tim Rocktäschel (2022). "Evolving curricula with regret-based environment design". *International Conference on Machine Learning*. PMLR, pp. 17473–17498 (cited on p. 29).

Patil, Shishir G., Tianjun Zhang, Vivian Fang, C. Noppapon, Roy Huang, Aaron Hao, Martin Casado, Joseph E. Gonzalez, Raluca Ada Popa & Ion Stoica (2024). "GoEX: Perspectives and Designs Towards a Runtime for Autonomous LLM Applications". *arXiv:2404.06921* (cited on p. 45).

Patil, Shishir G., Tianjun Zhang, Xin Wang & Joseph E. Gonzalez (2023). "Gorilla: Large Language Model Connected with Massive APIs". *ArXiv* abs/2305.15334 (cited on p. 24).

Paulin, James, Anisoara Calinescu & Michael Wooldridge (2019). "Understanding flash crash contagion and systemic risk: A micro–macro agent-based approach". *Journal of Economic Dynamics and Control* 100, pp. 200–229 (cited on p. 23).

Pavlova, Maya, Erik Brinkman, Krithika Iyer, Vitor Albiero, Joanna Bitton, Hailey Nguyen, Joe Li, Cristian Canton Ferrer, Ivan Evtimov & Aaron Grattafiori (2024). "Automated Red Teaming with GOAT: the Generative Offensive Agent Tester". *arXiv:2410.01606* (cited on p. 42).

Pearl, Judea (2009). *Causality*. Cambridge University Press (cited on p. 17).

Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese & Geoffrey Irving (2022a). "Red Teaming Language Models with Language Models". *arXiv:2202.03286* (cited on pp. 7, 18, 20, 42, 43).

Perez, Ethan, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer & Jared Kaplan (2022b). "Discovering Language Model Behaviors with Model-Written Evaluations". *arXiv:2212.09251* (cited on p. 29).

Perez, Fábio & Ian Ribeiro (2022). "Ignore Previous Prompt: Attack Techniques For Language Models". *ArXiv* abs/2211.09527 (cited on p. 52).

Perez, Jérémy, Corentin Léger, Marcela Ovando-Tellez, Chris Foulon, Joan Dussauld, Pierre-Yves Oudeyer & Clément Moulin-Frier (2024). "Cultural evolution in populations of Large Language Models". *arXiv:2403.08882* (cited on p. 27).

Peshkin, Leonid, Kee-Eung Kim, Nicolas Meuleau & Leslie Pack Kaelbling (2000). "Learning to Cooperate Via Policy Search". *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 489–496 (cited on pp. 10, 12, 15).

Peskov, Denis, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil & Jordan Boyd-Graber (2020). "It Takes Two to Lie: One to Lie, and One to Listen". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (cited on p. 23).

Piao, Jinghua, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li & James Evans (2025). "Emergence of human-like polarization among large language model agents". *arXiv:2501.05171* (cited on p. 47).

Piatti, Giorgio, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan & Rada Mihalcea (2024). "Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents". *arXiv:2404.16698* (cited on pp. 14, 16).

Piliouras, Georgios & Xiao Wang (2021). "Constants of Motion: The Antidote to Chaos in Optimization and Game Dynamics". *arXiv:2109.03974* (cited on p. 33).

Piliouras, Georgios & Fang-Yi Yu (2022). "Multi-agent Performative Prediction: From Global Stability and Optimality to Chaos". *arXiv:2201.10483* (cited on pp. 31, 32).

Poslad, Stefan, Patricia Charlton & Monique Calisti (2002). "Specifying standard security mechanisms in multi-agent systems". *Workshop on Deception, Fraud and Trust in Agent Societies*. Springer, pp. 163–176 (cited on p. 42).

Possajennikov, Alex (2000). "On the evolutionary stability of altruistic and spiteful preferences". *J. Econ. Behav. Organ.* 42.1, pp. 125–129 (cited on p. 28).

Powell, Robert (2006). "War as a Commitment Problem". *Int. Organ.* 60.1, pp. 169–203 (cited on p. 34).

Power, Alethea, Yuri Burda, Harri Edwards, Igor Babuschkin & Vedant Misra (2022). "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets". *arXiv:2201.02177* (cited on p. 32).

Prasad, Mahendra (2018). "Social Choice and the Value Alignment Problem". Ed. by Yampolskiy, Roman V. Chapman and Hall/CRC. Chap. 21, pp. 291–314 (cited on p. 47).

Prelec, Dražen (2004). "A Bayesian Truth Serum for Subjective Data". *Science* 306.5695, pp. 462–466 (cited on p. 22).

Qin, Yujia, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu & Maosong Sun (2023). "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs". *ArXiv* abs/2307.16789 (cited on p. 24).

Raghavan, Manish, Solon Barocas, Jon Kleinberg & Karen Levy (2020). "Mitigating bias in algorithmic hiring: evaluating claims and practices". *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (cited on p. 7).

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum & Michael Wellman (2019). "Machine behaviour". *Nature* 568.7753, pp. 477–486 (cited on p. 27).

Raman, Gururaghav, Jimmy Chih-Hsien Peng & Talal Rahwan (2019). "Manipulating Residents' Behavior to Attack the Urban Power Distribution System". *IEEE Transactions on Industrial Informatics* 15.10, pp. 5575–5587 (cited on p. 24).

Rand, David G., Corina E. Tarnita, Hisashi Ohtsuki & Martin A. Nowak (2013). "Evolution of fairness in the one-shot anonymous ultimatum game". *Proceedings of the National Academy of Sciences* 110.7, pp. 2581–2586 (cited on p. 30).

Rashid, Tabish, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster & Shimon Whiteson (2018). "QMIX: Monotonic Value Function Factorisation for Deep Multi-agent Reinforcement Learning". *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018.* Ed. by Dy, Jennifer G. & Krause, Andreas. Vol. 80, pp. 4292–4301 (cited on pp. 10, 12, 15).

Rees, Ray (1993). "Tacit Collusion". *Oxford Review of Economic Policy* 9.2, pp. 27–40 (cited on p. 17).

Reif, John H. (1984). "The complexity of two-player games of incomplete information". *Journal of Computer and System Sciences* 29.2, pp. 274–301 (cited on p. 11).

Renshaw, Jarrett & Trevor Hunnicutt (2024). "Biden, Xi agree that humans, not AI, should control nuclear arms". *Reuters* (cited on pp. 35, 46).

Reséndiz-Benhumea, Georgina Montserrat, Tom Froese, Gabriel Ramos-Fernández & Sandra E. Smith Aguilar (2019). "Applying Social Network Analysis to Agent-Based Models: A Case Study of Task Allocation in Swarm Robotics Inspired by Ant Foraging Behavior". *Artificial Life Conference Proceedings*, pp. 616–623 (cited on p. 26).

Reuel, Anka, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer & Robert Trager (2024a). "Open Problems in Technical AI Governance". *arXiv:2407.14981* (cited on pp. 7, 35, 44, 45).

Reuel, Anka, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy & Mykel Kochenderfer (2024b). "BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices". *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (cited on p. 45).

Reuel, Anka & Trond Arne Undheim (2024). "Generative AI Needs Adaptive Governance". *arXiv:2406.04554* (cited on p. 17).

Richardson, Lewis Fry (1960). *Arms and Insecurity: A Mathematical Study of the Causes and Origins of War*. Boxwood Press (cited on p. 31).

Richerson, Peter J. & Robert Boyd (2010). *Not By Genes Alone. How Culture Transformed Human Evolution*. University of Chicago Press (cited on p. 27).

Rismani, Shalaleh, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon & Negar Rostamzadeh (2023). "From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML". *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18 (cited on p. 45).

Rivera, Juan-Pablo, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith & Jacquelyn Schneider (2024). "Escalation Risks from Language Models in Military and Diplomatic Decision-Making". *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 836–898 (cited on p. 15).

Roemer, John E. (2010). "Kantian equilibrium". *Scandinavian Journal of Economics* 112.1, pp. 1–24 (cited on p. 17).

Roger, Fabien & Ryan Greenblatt (2023). "Preventing Language Models From Hiding Their Reasoning". *arXiv:2310.18512* (cited on pp. 18, 19).

Rusch, Hannes (2014). "The evolutionary interplay of intergroup conflict and altruism in humans: a review of parochial altruism theory and prospects for its extension". *Proc. Biol. Sci.* 281.1794, p. 20141539 (cited on p. 28).

Russell, Stuart (2019). *Human Compatible*. Penguin LCC US (cited on pp. 7, 44).

Sadeghi, McKenzie & Lorenzo Arvanitis (2023). "Rise of the Newsbots: AI-Generated News Websites Proliferating Online". *NewsGuard* (cited on p. 24).

Samvelyan, Mikayel, Akbir Khan, Michael Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Roberta Raileanu & Tim Rocktäschel (2023). "MAESTRO: Open-ended environment design for multi-agent reinforcement learning". *arXiv:2303.03376* (cited on p. 29).

Sandbrink, Jonas, Hamish Hobbs, Jacob Swett, Allan Dafoe & Anders Sandberg (2022). "Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks". *SSRN Electronic Journal* (cited on pp. 35, 36).

Sanders, James B. T., J. Doyne Farmer & Tobias Galla (2018). "The prevalence of chaotic dynamics in games with many players". *Scientific Reports* 8.1 (cited on pp. 5, 31–33).

Sandholm, Tuomas & Victor Lesser (2002). "Leveled-Commitment Contracting: A Backtracking Instrument for Multiagent Systems". *AI Magazine* 23.3, pp. 89–100 (cited on p. 35).

Sandholm, William H. (2010). *Population games and evolutionary dynamics*. MIT press (cited on p. 30).

Sangwan, Raghvinder S., Youakim Badr & Satish M. Srinivasan (2023). "Cybersecurity for AI Systems: A Survey". *Journal of Cybersecurity and Privacy* 3.2, pp. 166–190 (cited on p. 42).

Santos, Fernando P., Jorge M. Pacheco, Ana Paiva & Francisco C. Santos (2019). "Evolution of collective fairness in hybrid populations of humans and agents". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 6146–6153 (cited on p. 30).

Sastry, Shankar (1999). *Nonlinear Systems*. Springer New York (cited on p. 33).

Sato, Yuzuru, Eizo Akiyama & J. Doyne Farmer (2002). "Chaos in learning a simple two-person game". *Proceedings of the National Academy of Sciences* 99.7, pp. 4748–4751 (cited on p. 32).

Schäfer, Benjamin, Dirk Witthaut, Marc Timme & Vito Latora (2018). "Dynamically induced cascading failures in power grids". *Nature communications* 9.1, p. 1975 (cited on p. 33).

Schelling, Thomas C. (1980). *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press (cited on pp. 11, 16).

Schick, Timo, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda & Thomas Scialom (2023). "Toolformer: Language Models Can Teach

Themselves to Use Tools". *Thirty-seventh Conference on Neural Information Processing Systems* (cited on p. 37).

Schmeidler, David (1969). "The nucleolus of a characteristic function game". *SIAM Journal on applied mathematics* 17.6, pp. 1163–1170 (cited on p. 16).

Schmidt, Eric (2022). "AI, Great Power Competition & National Security". *Daedalus* 151, pp. 288–298 (cited on p. 26).

Schmitz, Patrick W. (2001). "The hold-up problem and incomplete contracts: a survey of recent topics in contract theory". *Bulletin of economic research* 53.1, pp. 1–17 (cited on p. 34).

Schneier, Bruce (2012). *Liars and Outliers: Enabling the Trust that Society Needs to Thrive*. 1st edition. Wiley (cited on p. 41).

Schroeder de Witt, Christian, Hawra Milani, Klaudia Krawiecka, Swapneel Mehta, Carla Cremer & Martin Strohmeier (2023a). *Multi-Agent Security Workshop at NeurIPS 2023* (cited on p. 39).

Schroeder de Witt, Christian, Samuel Sokota, J. Zico Kolter, Jakob Nicolaus Foerster & Martin Strohmeier (2023b). "Perfectly Secure Steganography Using Minimum Entropy Coupling". *The Eleventh International Conference on Learning Representations* (cited on pp. 18, 41).

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford & Oleg Klimov (2017). "Proximal Policy Optimization Algorithms". *arXiv:1707.06347* (cited on p. 52).

Schwettmann, Sarah, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau & Antonio Torralba (2023). "FIND: A Function Description Benchmark for Evaluating Interpretability Methods". *NeurIPS 2023* (cited on pp. 7, 20, 43).

SecureWorks (2023). *Unravelling the Attack Surface of AI Systems*. URL: https://www.secureworks.com/blog/unravelling-the-attack-surface-of-ai-systems (cited on p. 41).

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell & Abhishek Gupta (2023a). "Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives". *arXiv:2311.09227* (cited on p. 46).

Seger, Elizabeth, Aviv Ovadya, Divya Siddarth, Ben Garfinkel & Allan Dafoe (2023b). "Democratising AI: Multiple Meanings, Goals, and Methods". *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 715–722 (cited on p. 46).

Serapio-García, Greg, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust & Maja Matarić (2023). "Personality Traits in Large Language Models". *arXiv:2307.00184* (cited on p. 27).

Serrino, Jack, Max Kleiman-Weiner, David C. Parkes & Joshua B. Tenenbaum (2019). "Finding friend and foe in multi-agent games". *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc. (cited on p. 16).

Seth, Anil (2006). "Measuring emergence via nonlinear Granger causality". *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems* (cited on p. 38).

Shah, Nisarg (2023). "Pushing the Limits of Fairness in Algorithmic Decision-Making". *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7051–7056 (cited on p. 48).

Shakarian, Paulo, Jana Shakarian & Andrew Ruef (2013). "Chapter 12 - Can Cyber Warfare Leave a Nation in the Dark? Cyber Attacks Against Electrical Infrastructure". *Introduction to Cyber-Warfare*. Ed. by Shakarian, Paulo, Shakarian, Jana & Ruef, Andrew. Syngress, pp. 209–222 (cited on p. 23).

Shao, Yijia, Tianshi Li, Weiyan Shi, Yanchen Liu & Diyi Yang (2024). "PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action". *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (cited on p. 44).

Shapley, Lloyd S. (1953). "A value for n-person games". *Contribution to the Theory of Games* 2 (cited on p. 16).

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna M. Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang & Ethan Perez (2024). "Towards Understanding Sycophancy in Language Models". *The Twelfth International Conference on Learning Representations* (cited on p. 28).

Sharma, Prashant (2025). "How Cyberscammers Use AI to Manipulate Google Search Results". *Tech Pluto* (cited on p. 40).

Shavit, Yonadav (2023). "What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring". *ArXiv* abs/2303.11341 (cited on p. 36).

Shearer, Megan J., Gabriel Rauterberg & Michael P. Wellman (2023). "Learning to Manipulate a Financial Benchmark". *Fourth International Conference on Artificial Intelligence in Finance* (cited on p. 21).

Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia & Gurleen Virk (2023). "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction". *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723–741 (cited on pp. 8, 45).

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano & Allan Dafoe (2023). "Model evaluation for extreme risks". *arXiv:2305.15324* (cited on pp. 20, 29, 42, 45).

Shnayder, Victor, Arpit Agarwal, Rafael Frongillo & David C. Parkes (2016). "Informed Truthfulness in Multi-Task Peer Prediction". *Proceedings of the 2016 ACM Conference on Economics and Computation* (cited on p. 22).

Shoham, Yoav & Moshe Tennenholtz (1992). "On the Synthesis of Useful Social Laws for Artificial Agent Societies". *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 276–281 (cited on p. 12).

Shrivastava, Aryan, Jessica Hullman & Max Lamparth (2024). "Measuring Free-Form Decision-Making Inconsistency of Language Models in Military Crisis Simulations". *arXiv:2410.13204* (cited on p. 15).

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal (2024). "AI models collapse when trained on recursively generated data". *Nature* 631, pp. 755–759 (cited on pp. 26, 32).

Siegel, Zachary S., Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl & Arvind Narayanan (2024). "CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark". *arXiv:2409.11363* (cited on p. 45).

Siegenfeld, Alexander F. & Yaneer Bar-Yam (2020). "An introduction to complex systems science and its applications". *Complexity* 2020, pp. 1–16 (cited on p. 23).

Sigmund, Karl, Hannelore De Silva, Arne Traulsen & Christoph Hauert (2010). "Social learning promotes institutions for governing the commons". *Nature* 466.7308, pp. 861–863 (cited on p. 30).

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis (2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search". *Nature* 529.7587, pp. 484–489 (cited on pp. 15, 28).

Sinitsin, Anton, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov & Artem Babenko (2020). "Editable Neural Networks". *International Conference on Learning Representations* (cited on p. 20).

Siththaranjan, Anand, Cassidy Laidlaw & Dylan Hadfield-Menell (2024). "Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF". *The Twelfth International Conference on Learning Representations* (cited on p. 47).

Skitka, Linda J., Kathleen L. Mosier & Mark Burdick (1999). "Does automation bias decision-making?" *International Journal of Human-Computer Studies* 51.5, pp. 991–1006 (cited on p. 35).

Skopik, Florian & Timea Pahi (2020). "Under false flag: using technical artifacts for cyber attack attribution". *Cybersecurity* 3.1, p. 8 (cited on pp. 40, 41).

Slantchev, Branislav L. & Ahmer Tarar (2011). "Mutual optimism as a rationalist explanation of war". *American Journal of Political Science* 55.1, pp. 135–148 (cited on p. 21).

Small, Christopher T., Michael Bjorkegren, Timo Erkkilä, Lynette Shaw & Colin Megill (2021). "Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces". *RECERCA. Revista de Pensament i Anàlisi* 26.2, pp. 1–26 (cited on p. 46).

Small, Christopher T., Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli & Colin Megill (2023). "Opportunities and Risks of LLMs for Scalable Deliberation with Polis". *arXiv:2306.11932* (cited on pp. 13, 46).

Smith, J. Maynard & G. R. Price (1973). "The Logic of Animal Conflict". *Nature* 246.5427, pp. 15–18 (cited on p. 30).

Smith, John Maynard & Eörs Szathmáry (2020). *The Major Transitions in Evolution*. Previously issued in print: Oxford : W.H. Freeman/Spektrum, 1995; Oxford: Oxford University Press, 1997. - Includes bibliographical references and index. - Description based on print version record and publisher information. Oxford University Press (cited on p. 38).

Snell, Charlie, Jaehoon Lee, Kelvin Xu & Aviral Kumar (2024). "Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters". *arXiv:2408.03314* (cited on p. 47).

Snyder, Glenn H. (1971). """ Prisoner's Dilemma" and" Chicken" Models in International Politics". *International Studies Quarterly* 15.1, pp. 66–103 (cited on p. 13).

Solon, Olivia (2011). "How a book about flies came to be priced $24 million on Amazon". *Wired* (cited on p. 18).

Solum, Lawrence B. (1992). "Legal Personhood for Artificial Intelligences". *North Carolina Law Review* 70, pp. 1231–1290 (cited on pp. 17, 36, 46).

Sorensen, Taylor, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff & Yejin Choi (2024). "Position: a roadmap to pluralistic alignment". *Proceedings of the 41st International Conference on Machine Learning* (cited on p. 47).

Souly, Alexandra, Timon Willi, Akbir Khan, Robert Kirk, Chris Lu, Edward Grefenstette & Tim Rocktäschel (2023). "Leading the Pack: N-player Opponent Shaping". *arXiv:2312.12564* (cited on p. 16).

Sourbut, Oliver, Lewis Hammond & Harriet Wood (2024). "Cooperation and Control in Delegation Games". *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence* (cited on pp. 5, 24, 38, 43, 46).

Spohn, Wolfgang (2007). "Dependency equilibria". *Philosophy of Science* 74.5, pp. 775–789 (cited on p. 17).

Stadler, Theresa & Carmela Troncoso (2022). "Why the search for a privacy-preserving data sharing mechanism is failing". *Nature Computational Science* 2.4, pp. 208–210 (cited on p. 42).

Stanley, Kenneth O., Joel Lehman & Lisa Soros (2017). "Open-endedness: The last grand challenge you've never heard of". *O'Reilly Radar* (cited on p. 27).

Stastny, Julian, Maxime Riché, Alexander Lyzhov, Johannes Treutlein, Allan Dafoe & Jesse Clifton (2021). "Normative Disagreement as a Challenge for Cooperative AI". *arXiv:2111.13872* (cited on pp. 13, 16, 32).

Steimers, A. & Moritz Schneider (2022). “Sources of Risk of AI Systems”. *International Journal of Environmental Research and Public Health* 19 (cited on p. 26).

Stengel, Bernhard von & Shmuel Zamir (2010). “Leadership games with convex strategy sets”. *Games and Economic Behavior* 69.2, pp. 446–457 (cited on pp. 34, 47).

Stone, Peter, Gal Kaminka, Sarit Kraus & Jeffrey Rosenschein (2010). “Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination”. *Proceedings of the AAAI Conference on Artificial Intelligence* 24.1, pp. 1504–1509 (cited on pp. 10, 12, 15, 22, 29, 32).

Stroebl, Benedikt, Sayash Kapoor & Arvind Narayanan (2025). *HAL: A Holistic Agent Leaderboard for Centralized and Reproducible Agent Evaluation*. https://github.com/princeton-pli/hal-harness/ (cited on p. 45).

Suarez, Joseph, Yilun Du, Phillip Isola & Igor Mordatch (2019). “Neural MMO: A Massively Multiagent Game Environment for Training and Evaluating Intelligent Agents”. *arXiv:1903.00784* (cited on p. 38).

Subrahmanyam, Avanidhar (2013). “Algorithmic trading, the Flash Crash, and coordinated circuit breakers”. *Borsa Istanbul Review* 13.3, pp. 4–9 (cited on pp. 33, 46).

Sukhbaatar, Sainbayar, Arthur Szlam & Rob Fergus (2016). “Learning multiagent communication with backpropagation”. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2252–2260 (cited on pp. 18, 21).

Sun, Shuo, Rundong Wang & Bo An (2023a). “Reinforcement Learning for Quantitative Trading”. *ACM Trans. Intell. Syst. Technol.* 14.3 (cited on pp. 4, 5).

Sun, Xinyuan, Davide Crapis, Matt Stephenson, Barnabé Monnot, Thomas Thiery & Jonathan Passerat-Palmbach (2023b). “Cooperative AI via Decentralized Commitment Devices”. *arXiv:2311.07815* (cited on pp. 16, 42).

Sutton, Andrew & Reza Samavi (2018). “Tamper-Proof Privacy Auditing for Artificial Intelligence Systems”. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 5374–5378 (cited on p. 42).

Szabo, Claudia & Yong Meng Teo (2015). “Formalization of Weak Emergence in Multiagent Systems”. *ACM Transactions on Modeling and Computer Simulation* 26.1, pp. 1–25 (cited on p. 38).

Tadelis, Steven (2016). “Reputation and Feedback Systems in Online Platform Markets”. *Annual Review of Economics* 8.1, pp. 321–340 (cited on p. 36).

Talbot, David (2005). “Preventing ‘Fratricide’”. *MIT Technology Review* (cited on p. 35).

Talboy, Alaina N. & Elizabeth Fuller (2023). “Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption” (cited on p. 28).

Telikani, Akbar, Amirhessam Tahmassebi, Wolfgang Banzhaf & Amir H. Gandomi (2021). “Evolutionary machine learning: A survey”. *ACM Computing Surveys (CSUR)* 54.8, pp. 1–35 (cited on p. 27).

Tennenholtz, Moshe (2004). “Program equilibrium”. *Games and Economic Behavior* 49.2, pp. 363–373 (cited on pp. 16, 22, 34, 36).

Teo, Yong Meng, Ba Linh Luong & Claudia Szabo (2013). “Formalization of emergence in multi-agent systems”. *Proceedings of the 1st ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pp. 231–240 (cited on p. 38).

Terrucha, Inês, Elias Fernández Domingos, Pieter Simoens & Tom Lenaerts (2024). “Committing to the wrong artificial delegate in a collective-risk dilemma is better than directly committing mistakes”. *Scientific Reports* 14.1 (cited on p. 46).

Thurnherr, Lara, Robert Trager, Amin Oueslati, Christoph Winter, Cliodhna N’i Ghuidhir, Joe O’Brien, Jun Shern Chan, Lorenzo Pacchiardi, Anka Reuel, Merlin Stein, Oliver Guest, Oliver Sourbut, Renan Araujo, Seth Donoughe & Yi Zeng (2025). *Who Should Develop Which AI Evaluations?* Tech. rep. Oxford Martin School, University of Oxford (cited on p. 45).

Tian, Yu, Xiao Yang, Jingyuan Zhang, Yinpeng Dong & Hang Su (2023). “Evil Geniuses: Delving into the Safety of LLM-based Agents”. *arXiv:2311.11855* (cited on p. 26).

Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, Simon Staffell & José Jaime Villalobos (2023). "International Governance of Civilian AI: A Jurisdictional Certification Approach". *arXiv:2308.15514* (cited on pp. 7, 46).

Traulsen, Arne, Jorge M. Pacheco & Martin A. Nowak (2007). "Pairwise comparison and selection temperature in evolutionary game dynamics". *Journal of theoretical biology* 246.3, pp. 522–529 (cited on p. 30).

Traulsen, Arne, Dirk Semmann, Ralf D. Sommerfeld, Hans-Jürgen Krambeck & Manfred Milinski (2010). "Human strategy updating in evolutionary games". *Proceedings of the National Academy of Sciences* 107.7, pp. 2962–2966 (cited on p. 30).

Treutlein, Johannes, Michael Dennis, Caspar Oesterheld & Jakob Foerster (2021). "A New Formalism, Method and Open Issues for Zero-Shot Coordination". *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Meila, Marina & Zhang, Tong. Vol. 139, pp. 10413–10423 (cited on pp. 12, 16, 22).

Tucker, Mycal, Yilun Zhou & Julie Shah (2020). "Adversarially Guided Self-Play for Adopting Social Conventions". *arXiv:2001.05994* (cited on p. 23).

Turchin, Alexey & David Denkenberger (2018). "Classification of global catastrophic risks connected with artificial intelligence". *AI & SOCIETY* 35.1, pp. 147–163 (cited on p. 9).

Turlay, Emmanuel (n.d.). *What is Lineage Tracking in Machine Learning and why you need It*. Semantic. URL: https://www.sematic.dev/blog/what-is-lineage-tracking-in-machine-learning-and-why-you-need-it (cited on p. 42).

Turner, Alexander Matt & Prasad Tadepalli (2022). "Parametrically Retargetable Decision-Makers Tend To Seek Power". *Advances in Neural Information Processing Systems*. Ed. by Oh, Alice H., Agarwal, Alekh, Belgrave, Danielle & Cho, Kyunghyun (cited on p. 38).

Turner-Henderson, Tiffanie (2025). "Network Models: Triggering Marketing Network Effects With AI". *International Journal of Artificial Intelligence (AI) in Business and Management (IJAIBM)* 1.1, pp. 1–16 (cited on p. 26).

Tuyls, Karl & Ann Nowé (2005). "Evolutionary Game Theory and Multi-agent Reinforcement Learning". *The Knowledge Engineering Review* 20.1, pp. 63–90 (cited on p. 33).

U.S. Congress (2023). *Block Nuclear Launch by Autonomous Artificial Intelligence Act*. H.R.2894, 118th Congress (cited on pp. 17, 35, 46).

Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi & Sean Ekins (2022). "Dual use of artificial-intelligence-powered drug discovery". *Nature Machine Intelligence* 4.3, pp. 189–191 (cited on pp. 37, 44).

Uuk, Risto, Carlos Ignacio Gutierrez, Lode Lauwaert, Carina Prunkl & Lucia Velasco (2025). "A Taxonomy of Systemic Risks from General-Purpose AI". *SSRN Electronic Journal* (cited on p. 8).

Vallinder, Aron & Edward Hughes (2024). "Cultural Evolution of Cooperation among LLM Agents". *arXiv:2412.10270* (cited on pp. 28, 30).

Vallor, Shannon (2018). *Technology and the virtues. A philosophical guide to a future worth wanting*. First issued as an Oxford University Press paperback. Oxford University Press (cited on p. 28).

Van Loo, Rory (2019). "Digital Market Perfection". *Michigan Law Review* 117.5, p. 815 (cited on pp. 13, 46).

Vassilakopoulou, Polyxeni & Eli Hustad (2021). "Bridging Digital Divides: a Literature Review and Research Agenda for Information Systems Research". *Information Systems Frontiers* 25.3, pp. 955–969 (cited on p. 47).

Vegesna, Vinod Varma (2023). "Privacy-Preserving Techniques in AI-Powered Cyber Security: Challenges and Opportunities". *International Journal of Machine Learning for Sustainable Development* 5.4, pp. 1–8 (cited on pp. 36, 42).

Vestad, Arnstein & Bian Yang (2024). "A survey of agent-based modeling for cybersecurity". *Human Factors in Cybersecurity*. Vol. 127 (cited on pp. 26, 30, 42).

Vezhnevets, Alexander Sasha, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon & Joel Z. Leibo (2023). "Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia". *arXiv:2312.03664* (cited on pp. 26, 38).

Vié, Aymeric & Alfredo J. Morales (2021). "How connected is too connected? Impact of network topology on systemic risk and collapse of complex economic systems". *Computational economics* 57, pp. 1327–1351 (cited on p. 25).

Vinitsky, Eugene, Raphael Köster, John P. Agapiou, Edgar A. Duéñez-Guzmán, Alexander S. Vezhnevets & Joel Z. Leibo (2023). "A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings". *Collective Intelligence* 2.2, p. 263391372311620 (cited on p. 15).

Vlatakis-Gkaragkounis, Emmanouil-Vasileios, Lampros Flokas & Georgios Piliouras (2023). "Chaos persists in large-scale multi-agent learning despite adaptive learning rates". *arXiv:2306.01032* (cited on p. 32).

Wang, Ethan, Binghong Chen & Le Song (2021a). "Large Scale Coordination Transfer for Cooperative Multi-Agent Reinforcement Learning". *Deep RL Workshop at NeurIPS* (cited on p. 23).

Wang, Jane X., Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez-Guzmán & Joel Z. Leibo (2019a). "Evolving Intrinsic Motivations for Altruistic Behavior". *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 683–692 (cited on p. 16).

Wang, Jing, In Soo Ahn, Yufeng Lu & Tianyu Yang (2016). "A distributed detection algorithm for collective behaviors in multiagent systems". *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, pp. 372–377 (cited on p. 38).

Wang, Rui, Joel Lehman, Jeff Clune & Kenneth O. Stanley (2019b). "Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions". *arXiv:1901.01753* (cited on p. 29).

Wang, Tonghan, Paul Duetting, Dmitry Ivanov, Inbal Talgam-Cohen & David C. Parkes (2023). "Deep Contract Design via Discontinuous Networks". *Thirty-seventh Conference on Neural Information Processing Systems* (cited on p. 35).

Wang, Woodrow Z., Mark Beliaev, Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani & Dorsa Sadigh (2021b). "Emergent Prosociality in Multi-agent Games through Gifting". *arXiv:2105.06593* (cited on p. 15).

Wang, Xintong & Michael P. Wellman (2020). "Market manipulation: An adversarial learning framework for detection and evasion". *Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4626–4632 (cited on p. 21).

Wang, Zhen, Ruiqi Song, Chen Shen, Shiya Yin, Zhao Song, Balaraju Battu, Lei Shi, Danyang Jia, Talal Rahwan & Shuyue Hu (2024). "Large Language Models Overcome the Machine Penalty When Acting Fairly but Not When Acting Selfishly or Altruistically". *arXiv:2410.03724* (cited on p. 27).

Ward, Francis Rhys, Matt MacDermott, Francesco Belardinelli, Francesca Toni & Tom Everitt (2024). "The Reasons that Agents Act: Intention and Instrumental Goals". *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1901–1909 (cited on pp. 37, 38).

Ward, Francis Rhys, Francesca Toni, Francesco Belardinelli & Tom Everitt (2023). "Honesty Is the Best Policy: Defining and Mitigating AI Deception". *Thirty-seventh Conference on Neural Information Processing Systems* (cited on pp. 21, 23).

Watkins, Christopher J. C. H. & Peter Dayan (1992). "Q-learning". *Machine Learning* 8.3-4, pp. 279–292 (cited on p. 32).

Wei, Alexander, Nika Haghtalab & Jacob Steinhardt (2023). "Jailbroken: How Does LLM Safety Training Fail?" *Thirty-seventh Conference on Neural Information Processing Systems* (cited on pp. 41, 42).

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean & William Fedus (2022). "Emergent Abilities of Large Language Models". *Transactions on Machine Learning Research* (cited on p. 37).

Wei, Wenqi & Ling Liu (2024). "Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance". *ACM Comput. Surv.* (cited on pp. 41, 42, 44).

Weidinger, Laura, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield & Iason Gabriel (2023a). "Using the Veil of Ignorance to align AI systems with principles of justice". *Proceedings of the National Academy of Sciences* 120.18 (cited on p. 47).

Weidinger, Laura, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser & William Isaac (2023b). "Sociotechnical Safety Evaluation of Generative AI Systems". *arXiv:2310.11986* (cited on pp. 8, 43, 44).

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving & Iason Gabriel (2022). "Taxonomy of Risks posed by Language Models". *2022 ACM Conference on Fairness, Accountability, and Transparency* (cited on p. 8).

Werfel, Justin, Kirstin Petersen & Radhika Nagpal (2014). "Designing Collective Behavior in a Termite-Inspired Robot Construction Team". *Science* 343.6172, pp. 754–758 (cited on p. 37).

Werner, Tobias (2021). "Algorithmic and Human Collusion". *SSRN Electronic Journal* (cited on p. 19).

Wieting, Marcel & Geza Sapi (2021). "Algorithms in the Marketplace: An Empirical Analysis of Automated Pricing in E-Commerce". *SSRN Electronic Journal* 21-06 (cited on pp. 18, 19).

Willi, Timon, Alistair Letcher, Johannes Treutlein & Jakob N. Foerster (2022). "COLA: Consistent Learning with Opponent-Learning Awareness". *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Chaudhuri, Kamalika, Jegelka, Stefanie, Song, Le, Szepesvari, Csaba, Niu, Gang & Sabato, Sivan. Vol. 162, pp. 23804–23831 (cited on p. 16).

Wills, Peter (2024). "Care for Chatbots". *UBC Law Review (57:3)*. Forthcoming (cited on p. 46).

Witkowski, Jens & David C. Parkes (2012). "A robust Bayesian truth serum for small populations". *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp. 1492–1498 (cited on p. 22).

Wood, Elisabeth Jean (2008). "The social processes of civil war: The wartime transformation of social networks". *Annu. Rev. Polit. Sci.* 11, pp. 539–561 (cited on p. 23).

Wu, Feng, Lei Cui, Shaowen Yao & Shui Yu (2024a). "Inference Attacks: A Taxonomy, Survey, and Promising Directions". *arXiv:2406.02027* (cited on pp. 41, 42, 44).

Wu, Jibang, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan & Haifeng Xu (2022). "Sequential Information Design: Markov Persuasion Process and Its Efficient Reinforcement Learning". *arXiv:2202.10678* (cited on p. 22).

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger & Chi Wang (2024b). "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations". *First Conference on Language Modeling* (cited on p. 37).

Wylde, Allison (2021). "Zero trust: Never trust, always verify". *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–4 (cited on pp. 39, 41).

Xia, Haoxiang, Yanyan Du & Zhaoguo Xuan (2012). "Structural Evolution in Knowledge Transfer Network: An Agent-Based Model". *Workshop on Complex Networks* (cited on p. 26).

Xia, Lirong & Vincent Conitzer (2010). "Stackelberg Voting Games: Computational Aspects and Paradoxes". *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. Ed. by Fox, Maria & Poole, David, pp. 921–926 (cited on p. 35).

Xu, Zihao, Yi Liu, Gelei Deng, Yuekang Li & Stjepan Picek (2024). "A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models". *Findings of the Association for Computational Linguistics*, pp. 7432–7449 (cited on p. 25).

Yamin, Muhammad Mudassar, Mohib Ullah, Habib Ullah & Basel Katt (2021). "Weaponized AI for cyber attacks". *Journal of Information Security and Applications* 57, p. 102722 (cited on p. 15).

Yang, Jiachen, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes & Hongyuan Zha (2020). "Learning to Incentivize Other Learning Agents". *Adaptive and Learning Agents Workshop at AAMAS* (cited on p. 15).

Yang, Jiachen, Ethan Wang, Rakshit Trivedi, Tuo Zhao & Hongyuan Zha (2022). "Adaptive Incentive Design with Multi-Agent Meta-Gradient Reinforcement Learning". *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1436–1445 (cited on pp. 15, 33).

Yang, Kuan, Kejiang Chen, Weiming Zhang & Nenghai Yu (2019). "Provably Secure Generative Steganography Based on Autoregressive Model". *Digital Forensics and Watermarking*. Springer International Publishing, pp. 55–68 (cited on p. 18).

Yang, Kunhe & Hanrui Zhang (2024). "Computational Aspects of Bayesian Persuasion under Approximate Best Response". *arXiv:2402.07426* (cited on p. 22).

Yao, Andrew C. (1982). "Protocols for secure computations". *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*. ISSN: 0272-5428, pp. 160–164 (cited on pp. 36, 41).

Yoachimik, Omer & Jorge Pacheco (2024). *4.2 Tbps of bad packets and a whole lot more: Cloudflare's Q3 DDoS report*. Cloudflare. URL: https://blog.cloudflare.com/ddos-threat-report-for-2024-q3/ (cited on p. 39).

Yu, Miao, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Qingsong Wen, Kun Wang & Yang Wang (2024). "NetSafe: Exploring the Topological Safety of Multi-agent Networks". *arXiv:2410.15686* (cited on p. 26).

Yudkowsky, Eliezer (2008). "Artificial Intelligence as a positive and negative factor in global risk". *Global Catastrophic Risks*. Oxford University Press (cited on p. 43).

Zeeman, E. Christopher (1976). "Catastrophe Theory". *Scientific American* 234.4, pp. 65–83 (cited on pp. 31, 32).

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi & Cynthia Dwork (2013). "Learning fair representations". *International conference on machine learning*. PMLR, pp. 325–333 (cited on p. 48).

Zeng, Yi, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang & Bo Li (2024). "AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies". *arXiv:2406.17864* (cited on p. 8).

Zhang, Boyang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou & Yang Zhang (2024a). *Breaking Agents: Compromising Autonomous LLM Agents Through Malfunction Amplification* (cited on p. 41).

Zhang, Chongjie & Julie A Shah (2014a). "Fairness in Multi-Agent Sequential Decision-Making". *Advances in Neural Information Processing Systems*. Ed. by Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K.Q. Vol. 27 (cited on p. 7).

— (2014b). "Fairness in multi-agent sequential decision-making". *Advances in Neural Information Processing Systems* 27 (cited on p. 48).

Zhang, Haoqi & David Parkes (2008). "Value-based policy teaching with active indirect elicitation". *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1*, pp. 208–214 (cited on pp. 15, 33).

Zhang, Kaiqing, Sham Kakade, Tamer Basar & Lin Yang (2020). "Model-Based Multi-Agent RL in Zero-Sum Markov Games with Near-Optimal Sample Complexity". *Advances in Neural Information Processing Systems*. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H. Vol. 33, pp. 1166–1178 (cited on p. 15).

Zhang, Kaiqing, Zhuoran Yang, Han Liu, Tong Zhang & Tamer Basar (2018). "Fully Decentralized Multi-agent Reinforcement Learning with Networked Agents". *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Dy, Jennifer & Krause, Andreas. Vol. 80, pp. 5872–5881 (cited on p. 21).

Zhang, Yadong, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan & Furu Wei (2024b). "LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models". *arXiv:2404.01230* (cited on p. 30).

Zhao, Mingyi, Aron Laszka & Jens Grossklags (2017). "Devising Effective Policies for Bug-Bounty Platforms and Security Vulnerability Discovery". *Journal of Information Policy* 7, pp. 372–418 (cited on p. 44).

Zhao, Ying & Jinjun Chen (2022). "A survey on differential privacy for unstructured data content". *ACM Computing Surveys (CSUR)* 54.10s, pp. 1–28 (cited on p. 48).

Zheng, Stephan, Alexander Trott, Sunil Srinivasa, David C. Parkes & Richard Socher (2022). "The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning". *Science Advances* 8.18 (cited on pp. 15, 33, 38).

Zhi-Xuan, Tan (2022). *What Should AI Owe To Us? Accountable and Aligned AI Systems via Contractualist AI Alignment*. Alignment Forum. URL: https://www.alignmentforum.org/posts/Cty2rSMut483QgBQ2 (cited on p. 47).

Zhi-Xuan, Tan, Micah Carroll, Matija Franklin & Hal Ashton (2024). "Beyond Preferences in AI Alignment". *Philosophical Studies* (cited on p. 47).

Zhou, Jiawei, Yixuan Zhang, Qianni Luo, Andrea G. Parker & Munmun De Choudhury (2023). "Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions". *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20 (cited on pp. 21, 47).

Zhu, Hao, Graham Neubig & Yonatan Bisk (2021). "Few-shot Language Coordination by Modeling Theory of Mind". *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Meila, Marina & Zhang, Tong. Vol. 139, pp. 12901–12911 (cited on pp. 12, 22, 23).

Zhu, Shuhui, Baoxiang Wang, Sriram Ganapathi Subramanian & Pascal Poupart (2025). "Learning to Negotiate via Voluntary Commitment". *The 28th International Conference on Artificial Intelligence and Statistics* (cited on p. 35).

Ziegler, Daniel M., Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, Buck Shlegeris & Nate Thomas (2022). "Adversarial training for high-stakes reliability". *NeurIPS* (cited on pp. 18, 20, 43).

Zimmaro, Filippo, Manuel Miranda, José María Ramos Fernández, Jesús A. Moreno López, Max Reddel, Valeria Widler, Alberto Antonioni & The Anh Han (2024). "Emergence of cooperation in the one-shot Prisoner's dilemma through Discriminatory and Samaritan AIs". *Journal of the Royal Society Interface* 21.218, p. 20240212 (cited on pp. 27, 30).

Zinkevich, Martin, Amy Greenwald & Michael L. Littman (2005). "Cyclic Equilibria in Markov Games". *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pp. 1641–1648 (cited on p. 32).

Ziyin, Liu & Masahito Ueda (2022). "Exact Phase Transitions in Deep Learning". *arXiv:2205.12510* (cited on p. 32).

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter & Matt Fredrikson (2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models* (cited on p. 15).

Zweetsloot, Remco & Allan Dafoe (2019). *Thinking about Risks from AI: Accidents, Misuse and Structure*. Lawfare Blog (cited on p. 44).