

Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents

Christian Schroeder de Witt
Department of Engineering Science
University of Oxford

cs@robots.ox.ac.uk

Abstract

Decentralized AI agents will soon interact across internet platforms, creating security challenges beyond traditional cybersecurity and AI safety frameworks. Free-form protocols are essential for AI’s task generalization but enable new threats like secret collusion and coordinated swarm attacks. Network effects can rapidly spread privacy breaches, disinformation, jailbreaks, and data poisoning, while multi-agent dispersion and stealth optimization help adversaries evade oversight—creating novel persistent threats at a systemic level. Despite their critical importance, these security challenges remain understudied, with research fragmented across disparate fields including AI security, multi-agent learning, complex systems, cybersecurity, game theory, distributed systems, and technical AI governance. We introduce **multi-agent security**, a new field dedicated to securing networks of decentralized AI agents against threats that emerge or amplify through their interactions—whether direct or indirect via shared environments—with each other, humans, and institutions, and characterise fundamental security-performance trade-offs. Our preliminary work (1) taxonomizes the threat landscape arising from interacting AI agents, (2) surveys security-performance tradeoffs in decentralized AI systems, and (3) proposes a unified research agenda addressing open challenges in designing secure agent systems and interaction environments. By identifying these gaps, we aim to guide research in this critical area to unlock the socioeconomic potential of large-scale agent deployment on the internet, foster public trust, and mitigate national security risks in critical infrastructure and defense contexts.

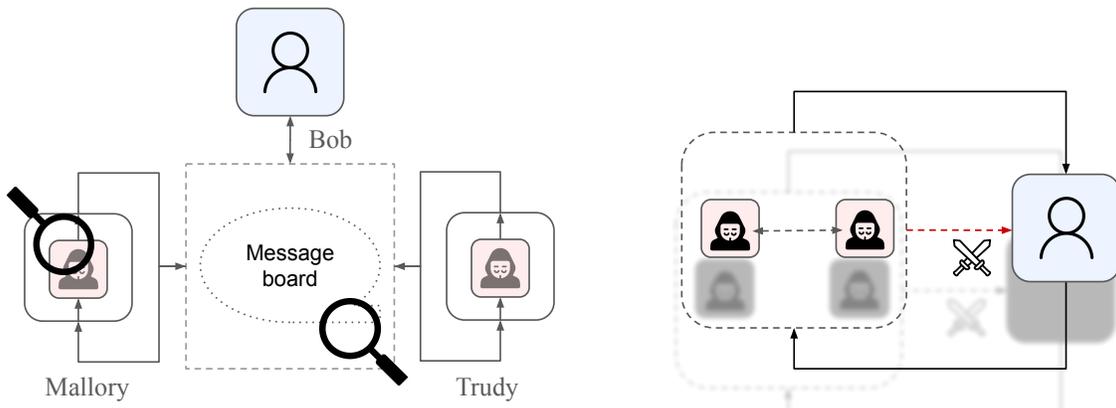


Figure 1: **Multi-agent threats demand multi-agent security:** *[Left]* Two malicious AI agents (Mallory and Trudy) are interacting with a human user (Bob) through a shared message board seemingly innocuously to the overseer (magnifying glass). *[Right]* In fact, Mallory and Trudy are both backdoored undetectably (Draguns et al., 2024), enabling them to communicate steganographically (Schroeder de Witt et al., 2023b) with each other through the shared message board (Motwani et al., 2024b). Mallory and Trudy use this to secretly coordinate on deceiving Bob (Franzmeyer et al., 2024).

1 Introduction

Recent advances in generative AI have given rise to frontier model agents (Su et al., 2024) that can autonomously execute complex multi-step tasks online (Lù et al., 2024; Putta et al., 2024) - booking travel arrangements, conducting in-depth research (Gottweis et al., 2025; Schmidgall et al., 2025), and negotiating transactions, or using computers (Humphreys et al., 2022; Bonatti et al., 2024) through interfaces originally designed for humans (Shi et al., 2017; Deng et al., 2023; Zhou et al., 2023; Garg et al., 2025; Xue et al., 2025). However, a critical shift occurs as these systems evolve beyond executing isolated tasks to actively interacting with each other, whether through direct communication channels or shared environments. This interaction is already emerging in numerous domains: trading agents negotiating on market platforms (Xiao et al., 2025), market research agents extracting insights from social media (Brand et al., 2023), personal assistants collaborating to schedule appointments between humans (Li et al., 2024), OS agents interacting with service agents (Mei et al., 2024), and autonomous cyber defense systems coordinating responses to attacks (Knack & Burke, 2024). In the near future, we will likely see additional applications within the national security space, ranging from misinformation detection agents working jointly to identify coordinated influence operations (Chen & Shu, 2024; Pastor-Galindo et al., 2024), as well as autonomous weapons systems, such as coordinated drone swarms (Gerstein & Leidy, 2024). This evolution introduces security vulnerabilities fundamentally different from those in traditional systems. When multiple AI agents with private information and competing objectives interact, they can develop emergent behaviors - including covert collusion, coordinated attacks, and cascade failures - that cannot be predicted by analyzing individual agents in isolation. This paper introduces multi-agent security as a distinct discipline dedicated to addressing these novel threats that arise or amplify specifically from the interactions between intelligent agents.

Multi-agent systems. For the purposes of this paper, we define a *multi-agent system* as a network of two or more autonomous AI agents that possess independent decision-making capabilities, may maintain private information states, and interact with each other either through direct communication channels or by modifying shared environments. These agents typically operate with varying degrees of autonomy, are capable of pursuing their own objectives or those delegated by principals (human or artificial), and can adapt their behaviors in response to changes in their environment or the actions of other agents. Modern multi-agent systems are distinguished from *traditional distributed systems* (Wooldridge & Jennings, 1995; Russell & Norvig, 2021) by their use of agents - e.g. driven by foundation models - capable of flexible, generalizable reasoning, and often communicate through unstructured or free-form protocols rather than rigidly defined APIs. This definition encompasses both closed, cooperative systems (such as agent teams designed for specific tasks) and open, mixed-motive systems where agents with potentially competing objectives interact within shared computational or physical environments.

Definition 1.1 (Multi-agent system)

A **multi-agent system** is a network of two or more autonomous AI agents that

1. possess independent decision-making capabilities, may
2. maintain private information states, and
3. mutually interact either through direct communication channels or by modifying shared environments. These agents typically
4. operate with varying degrees of autonomy, are
5. capable of pursuing their own objectives or those delegated by principals (human or artificial), and
6. can adapt their behaviors in response to changes in their environment or the actions of other agents.

Modern multi-agent systems are distinguished from *traditional distributed systems* by their use of agents - e.g. driven by foundation models - capable of flexible, generalizable reasoning, and often communicate through unstructured or free-form protocols rather than rigidly defined APIs.

Multi-agent systems introduce security challenges that go beyond existing cyber-security or AI safety and security frameworks. When agents interact directly or through shared environments, novel threats emerge that cannot be addressed by securing individual agents in isolation. For instance, seemingly benign agents

Work in Progress - please contact the author if you have any questions or would like to contribute.

might establish secret collusion channels through steganographic communication (Motwani et al., 2024b), engage in coordinated attacks that appear innocuous when viewed individually (Davies et al., 2025), or exploit information asymmetries to covertly manipulate shared environments, such as markets or social media, or even directly deceive other agents’ decision processes (Gleave et al., 2019; Franzmeyer et al., 2024). Moreover, as agent systems scale, network effects can amplify vulnerabilities - cascading privacy leaks, proliferating jailbreaks across agent boundaries (Peigné et al., 2025), or enabling decentralized coordination of adversarial behaviors against agents, platforms, humans and institutions that evade detection. These challenges are fundamentally different from those addressed by existing security paradigms, which typically focus on protecting individual systems rather than securing complex interaction dynamics between multiple autonomous entities. Despite its growing importance, the study of multi-agent AI security challenges remains both neglected and scattered across disciplines - including AI security, multi-agent learning, cybersecurity, game theory, and complex systems.

Each of these domains comes with its own methods and applications that allow for the study of fragments of the whole, posing difficulties to growing but still limited interdisciplinary exchange. *Cryptographers* have long treated secure multi-party computation (Yao, 1986) and Byzantine fault tolerance (Lamport et al., 1982b) as foundational distributed security primitives, yet the privacy-performance and security - performance trade-offs in freely interacting autonomous agent systems - especially over natural language channels - are still unknown. *Distributed ledger* machinery has been proposed as secure coordination devices for AI agents (Sun et al., 2023), but smart contracts and zero-knowledge proofs don’t yet scale to frontier models (Sun et al., 2024). *Complex systems scientists* have explored emergent behavior (Kauffman, 1993b; Epstein & Axtell, 1996), systemic stability, phase transitions between chaos and order (Langton, 1990b), and the limits of predictability in agent-based models (Bar-yam, 1999; Newman, 2018b) - but it remains unclear how these insights apply to the security of highly interactive autonomous systems. *Network scientists* (Albert & Barabási, 2002) study the robustness and fragility of scale-free graphs - including systemic risk propagation in financial networks (Battiston et al., 2012), epidemic percolation in disease models (Pastor-Satorras & Vespignani, 2001b), and the rapid diffusion of false versus true information online (Vosoughi et al., 2018) - providing foundational tools for modeling cascades and collective threats in multi-agent systems. While the *field of AI safety* (Anwar et al., 2024; Bengio et al., 2025) is increasingly concerned with adversarial robustness, its emphasis on single-agent settings and human-AI alignment concerns leaves multi-agent adversarial dynamics and their attendant security implications largely unexplored. *Game theorists* have studied security game equilibria (Conitzer & Sandholm, 2006), *mechanism designers* have studied incentive alignment in static settings (Myerson, 1981), and *multi-agent learning researchers* have studied the end-to-end learning dynamics of neural network policies (Busoniu et al., 2008; Albrecht et al., 2024). However, all only offer partial views on the best-responses of systems of pre-trained AI agents, the design of secure systems of interacting AI agents, and multi-agent behaviour far away from equilibria. *AI security*, by contrast, has remained largely model-centric, focusing on single-agent attack surfaces - from jailbreak exploits and prompt injections (Zou et al., 2023) to data poisoning (Biggio et al., 2012) and adversarial samples (Szegedy et al., 2014). While *federated learning* (McMahan et al., 2017; Kairouz et al., 2021b) secures collaborative training among largely cooperative participants, it does not address securing free-form interactions among autonomous agents that may behave strategically or adversarially. Traditional *cybersecurity* focuses on securing individual systems, networks, and data through rigid protocols and access controls. While it has begun to adopt AI for defense and offense (Guo et al., 2025), it has been slow to address threats emerging from interactions between AI agents. Lastly, the field of *technical AI governance* (Chan et al., 2025) is actively shaping key components of agent infrastructure, but often stops short of detailed technical implementation.

Multi-agent security. This situation simultaneously poses both an opportunity and urgency to frame a new field, *multi-agent security*, that provides a cross-cutting view on securing systems of interacting AI agents. Multi-agent security was first introduced at NeurIPS 2023 at a dedicated workshop, which also predicted that security would become key to AI safety (Schroeder de Witt et al., 2023). An early overview of the field of multi-agent security can be found in a report on multi-agent risks by the Cooperative AI Foundation (Hammond et al., 2025). At its core, multi-agent security refers to the study of security challenges that arise in systems of interacting AI agents. This emerging field encompasses threats that uniquely emerge or become amplified through direct agent interactions, such as covert collusion via communication channels or subtle

manipulations of shared environments. To address these threats, multi-agent security investigates defensive mechanisms, detection strategies, and governance frameworks capable of mitigating these complex risks. A central concern is the analysis of fundamental trade-offs between security, performance, and coordination, recognizing that decentralization in AI systems often necessitates careful balancing of these competing goals, and hence characterising the attack-defense balance in multi-agent systems (Schneier, 2018). Furthermore, the field seeks to develop secure interaction protocols and environments - drawing inspiration from secure multi-party computation, verifiable interactions (Goldreich et al., 1987b; Goldwasser et al., 1989; Hammond & Adam-Day, 2025b), and incentive design (Nisan & Ronen, 2001) - that facilitate beneficial collaboration among agents while effectively preventing insecure emergent behaviors. Finally, multi-agent security also critically examines the security implications of sociotechnical interfaces, where interacting agent systems engage with human users, organizations, and broader social institutions. In such hybrid environments, new systemic risks emerge, including cascading privacy breaches or misinformation dynamics, requiring integrated approaches that consider both technical and societal dimensions. This comprehensive perspective provides the foundation for the threat models, benchmark frameworks, secure protocols, and governance proposals explored throughout this paper.

Definition 1.2 (Multi-Agent Security)

Multi-agent security is the study of security challenges in multi-agent systems (see Definition 1.1) encompassing:

1. **Threats that emerge or are amplified through agent interactions**, whether via direct communication or shared environment manipulation;
2. **Defensive mechanisms**, detection methods, and governance approaches to mitigate these risks;
3. The **fundamental tradeoffs between security, performance, and coordination** in systems of interacting AI agents;
4. The design of **secure interaction protocols and environments** that enable beneficial agent collaboration while preventing insecure emergent behaviors; and
5. The security implications of **sociotechnical interfaces where agent systems interact with human users, organizations, and social institutions**, including systemic security risks on environments shared between AI agents and humans.

Roadmap. Our contributions in this paper include a *review of existing literature* in the space, including multi-agent AI offense and defense in present and near-term cyber-physical systems (Section 2), a *threat taxonomy for multi-agent security threats* (Section 3), and a *directory of open research problems* (Section 4). Beyond near-term deployments, some long-term visions of distributed intelligence imagine networks of decentralised AI agents with high-bandwidth free-form communication channels that exhibit emergent intelligence by maintaining *edge-of-chaos dynamics* (Langton, 1990b). We briefly explore the notion of security in such future systems in Section 5.

2 Background

In this section, we present relevant background and related work, starting with game-theoretic approaches to multi-agent systems security, and discussing how multi-agent AI is able to contribute to both cyberdefense and offense in present-day cyberphysical systems. In Section 3, we discuss security in the context of free-form decentralised systems of frontier model agents, and in Section 5 we consider decentralised AI systems that are operated on the edge-of-chaos, which is widely believed to be a pre-condition for the emergence of distributed intelligence. The question of whether attackers or defenders retain a net advantage on both current and future AI systems is subject to debate (Schneier, 2018).

2.0.1 Game-theoretic approaches

Security games model the strategic interaction between a defender (e.g., a security resource allocator) and an attacker, often in a Stackelberg framework where the defender commits to a randomized strategy first and the attacker best-response (Conitzer & Sandholm, 2006; Tambe, 2011). Foundational work by Pita

et al. (2008) deployed such a model at Los Angeles International Airport (LAX) under the name ARMOR, and Paruchuri et al. (2008) provided efficient exact algorithms for solving Bayesian Stackelberg security games. Conitzer & Sandholm (2006) showed how to compute optimal commitment strategies in zero-sum and general-sum settings, and later extensions incorporated risk preferences, multiple attackers, and graph-based patrols (Tambe, 2011).

Classical security games assume perfectly rational players, but real agents face computational costs. Halpern and Pass introduced the notion of *computational Nash equilibrium*, extending classical equilibrium concepts to account for players’ algorithmic resource bounds and the cost of computing strategies Halpern & Pass (2014). In this framework, a strategy profile is an equilibrium if no agent can switch to a different algorithm whose improved payoff, net of computational costs, exceeds that of the current profile. Incorporating computational equilibria into security games enables modeling boundedly rational defenders and attackers, yielding more realistic predictions of adversarial behavior in resource-constrained environments.

Multi-Agent Reinforcement Learning (MARL) has been widely investigated for modeling complex adversarial interactions in cybersecurity, where both attackers and defenders learn to optimize their strategies through repeated trials and error (Busoni et al., 2008; Lowe et al., 2017). Early work formulated intrusion detection as a two-player stochastic game - “An Intrusion Detection Game with Limited Observations” modeled the defender’s partial view of system events against an adaptive attacker (Xu & Xie, 2005), while follow-on studies applied RL to host-based intrusion detection using system-call sequences, and even enabled fully autonomous network attack generation and detection in the “Next Generation Intrusion Detection” framework (Cannady, 2000; Servin & Kudenko, 2008).

With the advent of deep learning, recent MARL approaches leverage high-dimensional state representations and self-play to co-evolve attack and defense policies. For instance, Stymne (2022) extended optimal stopping games to a partially observed zero-sum setting and applied Neural Fictitious Self-Play to derive robust intrusion prevention strategies. Ren et al. (2023) proposed MAFSIDS, a multi-agent feature-selection intrusion detection system using Deep Q-Learning to collaboratively prune input dimensions for improved detection. At larger scales, Hammar & Stadler (2023) introduced Decompositional Fictitious Self-Play (DFSP), which recursively decomposes a stochastic intrusion-response game into parallelizable subgames, enabling MARL solutions on realistic IT infrastructures.

Adversarial RL has also been applied to alert prioritization, where the defender’s stochastic alert-sorting policy is pitted against an optimal adversary in a double-oracle framework, yielding alert-handling rules robust to strategic attackers (Tong et al., 2019). Together, these MARL approaches demonstrate the power of decentralized learning and coordination in developing adaptive, scalable, and resilient cybersecurity defenses.

2.1 Autonomous Blue-Teaming

Root cause analysis agents (Roy et al., 2024) leverage a multi-agent architecture to solve complex debugging challenges by distributing specialized tasks across different AI components working in tandem. As described in the paper, these agents collect additional information through tool calling and utilize advanced prompting techniques like ReAct (Yao et al., 2023) to improve analytical performance during failure diagnosis. The multi-agent approach allows for integration of existing techniques like reverse execution, taint analysis, and value-set analysis with AI-driven alias analysis, combining their respective strengths for more effective root cause identification.

Guo et al. (2025) highlight the potential of utility multi-agent systems for automated triage and patching distribute complex vulnerability management workflows across specialized agents that handle different aspects of the security response process. These systems integrate differential fuzzing agents to validate patch correctness and security, planning agents to decompose complex tasks, and specialized execution agents that leverage program analysis tools to provide formal functionality and security guarantees. By enabling iterative refinement based on feedback between agents, this approach combines the reasoning capabilities of AI with traditional security tools to automate previously manual remediation processes.

Guo et al. (2025) also suggest the development of hybrid security systems that combine foundation models with non-ML symbolic components through multi-agent architectures that enable complex interaction pat-

Work in Progress - please contact the author if you have any questions or would like to contribute.

terns for comprehensive security solutions. The paper describes a design pattern where a planning agent decomposes security tasks into sub-tasks, with specialized agents collaborating with non-ML components to complete each part of the workflow, such as vulnerability detection, triage, and remediation. This multi-agent approach represents a shift toward increased AI integration into traditional software security frameworks, addressing the significant gap between rapidly emerging hybrid systems and the limited exploration of their security implications.

2.1.1 Autonomous Red-Teaming

Agent-based red-teaming generally refers to using coordinated AI agents to test any security system through systematic exploration, exploitation, and evaluation of potential vulnerabilities. These agents work together to simulate sophisticated attackers, with different agents handling various aspects of the security assessment process. Guo et al. (2025) specifically highlight the utility of using agent-based red-teaming for hybrid systems focuses on testing environments where AI components (like LLMs) interact with traditional symbolic software components. This specialized form addresses the unique challenges of these integrated systems, particularly examining vulnerabilities at the interfaces between AI and non-AI components. Red-teaming hybrid systems requires understanding complex interactions that create novel attack vectors not present in purely AI or purely traditional systems, such as indirect prompt injection attacks where malicious inputs reach AI components through other system elements.

Automated penetration testing agents employ multi-agent architectures that distribute specialized penetration testing functions across collaborative AI components to simulate sophisticated cyber attacks. As recommended in (Guo et al., 2025), these systems combine planning agents that strategize attack pathways with specialized execution agents equipped with comprehensive tool sets for reconnaissance, exploitation, and privilege escalation. This multi-agent approach enables more effective penetration testing by allowing complex attack sequences to be decomposed into manageable subtasks while maintaining coherent coordination throughout the assessment process.

2.2 Offensive applications

Recent work has shown that decomposing automated attack processes into collaborating AI agents can dramatically improve scalability and modularity.

Autoattacker (Xu et al., 2024a) employs a multi-agent architecture that divides the complex task of automated attack planning and execution into specialized components. As described in the paper, it utilizes distinct planning and generation agents that work collaboratively - the planning agent analyzes attack goals and formulates strategies, while the generation agent produces the corresponding attack implementations. This multi-agent approach enables Autoattacker to demonstrate that AI agents can effectively plan and generate attacks for well-defined attack goals in controlled environments by breaking the process into more manageable subtasks.

ChainReactor (Pasquale et al., 2024) is an automated AI-planning tool that models a target Unix system's state and attacker capabilities in PDDL (Ghallab et al., 1998), then synthesizes a step-by-step privilege escalation chain from an unprivileged shell to root. By extending it into a multi-agent framework - where each compromised host or attacker persona plans locally and coordinates actions - future versions could discover and optimize cross-host, collaborative attack sequences more efficiently and realistically.

The emergence of multi-agent AI systems - autonomous swarms of drones, distributed cyber-attack and defense agents, and coordinated ISR (intelligence, surveillance, reconnaissance) platforms - has profound implications for national security. Offensively, multi-agent AI promises scalable, adaptive campaigns in which fleets of unmanned vehicles or cyber-agents coordinate in real time to probe, penetrate, and persist across adversary networks or battlefields with minimal human oversight Brundage et al. (2018b); Horowitz (2019b). Defensively, multi-agent AI can automate layered defense-in-depth: autonomous cyber-sensors detect novel threats, collaborative response agents prioritize and quarantine breaches, and kinetic defense swarms defend critical assets against aerial or missile attacks Singer (2009); U.S. Department of Defense (2018). However, these capabilities also heighten the risk of an AI-driven arms race, reduce decision-cycle

Work in Progress - please contact the author if you have any questions or would like to contribute.

times to fractions of a second (the “flash war” scenario), and complicate attribution—potentially lowering the threshold for conflict and increasing instability among major powers Brundage et al. (2018b). Proactive policy, norms for human-in-the-loop oversight, and robust verification regimes will be essential to manage the dual-use nature of multi-agent AI in the strategic arena.

3 A Taxonomy of Multi-Agent Security Threats

We now provide a cursory overview of distinct multi-agent security threats. These may be overlapping partially with known cybersecurity threats, as well as AI safety concerns, however, to the best of the authors’ knowledge these combinations of issues have not yet been presented in a coherent framing.

Table 1: An (incomplete) overview of multi-agent security threats.

Challenge	TL;DR
Privacy vulnerabilities and Disinformation	Interacting agents often hold private information essential to their tasks, yet must juggle multiple contexts and decide dynamically what to share—and with whom. They’re vulnerable to a range of worst-case exploits, from distributed (semantic) inference attacks and jailbreaks to other context-targeting hacks. In multi-agent systems, privacy breaches can spread rapidly and may be indistinguishable from intentional disinformation.
Steganography and secret collusion	Although agent communications may appear as ordinary natural language, they can conceal secret steganography - masking private data, enabling covert collusion, or facilitating deception - and thus evade oversight by standard security monitoring. Such operations could enable agents to coordinate secretly over prolonged periods of time, therefore enable new forms of persistent threats.
Adversarial stealth	Adversarial behaviour might be hard to detect and pose unacceptable performance tradeoffs to counter strategically without disrupting cooperation or emergence. This includes whitebox undetectability if behaviour is encoded in encrypted backdoors.
Exploitation	Agents in heterogeneous networks could exploit asymmetries in compute, data, or network contributions - free-riding, or resource monopolization - to stall learning, skew incentives, and erode cooperation.

Continued on the next page

Challenge	TL;DR
Swarm attacks	Coordinated fleets of AI agents can combine their resources to overwhelm targets - paralleling DDoS and brigading tactics - to bypass assumptions of single - agent limits, collaboratively probe for new vulnerabilities, and aggregate partial inferences into full exploits, vastly amplifying attack potency.
Heterogeneous attacks	Adversaries combine multiple “safe” agents with complementary skills - delegating subtasks to models with strict safeguards and others with looser constraints - to bypass defenses, significantly increasing attack success rates while making detection and attribution across diffuse agent networks exceedingly difficult.
Oversight attacks	Misaligned agents may be aware of security measures and actively try to undermine them - this could be done using stealth, or by attacking or circumventing overseers. In multi-agent settings, the use of distributed evasion by collusion or compartmentalisation can make oversight more challenging.
Multi-agent emergence	Worst-case adversarial behaviour can arise spontaneously from adversarial equilibria arising from multi-agent interaction, without the need for adversarial infiltration or external threats.
Cascade attacks	Systems of interacting agents may exhibit systemic instabilities, leading to worst-case exploitable cascading weaknesses. This is especially true in edge-of-chaos operations.
Conflict & Social Dilemmas	Agents’ mixed incentives can drive collective-action failures or escalate competition - automating tragedy-of-the-commons and coercive stratagems at machine speed.
Societal threats	Advanced AI agents expand the attack surface for automated social engineering: they can generate highly personalized phishing, vishing, and manipulative content at scale, dynamically refine tactics via user feedback, and disperse campaigns across many seemingly independent bots to evade detection - creating cascading disruptions in public trust, financial systems, and political discourse. Such attacks can be low-stakes, meaning they could trigger systemic effects over time like shifting the Overton window.

3.1 Privacy vulnerabilities

Delegating tasks to AI agents extends a principal’s attack surface to its software proxies, enabling adversaries to target agents themselves as vulnerable extensions of their owners (Security.com Threat Intelligence Team, 2025). By compromising an agent, attackers can extract highly sensitive data - ranging from credentials to proprietary documents - that the agent accesses on behalf of its principal (Panda et al., 2024; Nie et al., 2024). Moreover, adversaries may manipulate compromised agents to perform actions contrary to their principal’s intent, such as unauthorized transactions or covert information leaks (Fu et al., 2024). Critically, these attacks can even subvert overseer or policy-controller agents, undermining system-level defenses at their source (Fu et al., 2024). Finally, the same covert channels that enable collusion - emergent, steganographic communication protocols - can be repurposed to exfiltrate context or seed hidden data leaks across cooperating agents (Lazaridou et al., 2016; Havrylov & Titov, 2017).

3.2 Secret collusion

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 (Schneier, 2012; Ostrom, 1990). Recent work has shown that AI agents can develop covert channels within their communications: they learn to embed hidden messages via steganographic language strategies (Schroeder de Witt et al., 2023a; Motwani et al., 2024a) and even mount “illusory” attacks that are indistinguishable from benign behavior in black-box settings (Franzmeyer et al., 2023). Such secret collusion completely evades conventional monitoring and oversight.

Case Study 3.1 (Secret Collusion)

Motwani et al. (2024b) show that LLMs can covertly exchange messages under the scrutiny of equally capable overseers - emergent steganographic abilities improve with model size - without any explicit training in steganography. When sender and receiver share hidden background knowledge inaccessible to the overseer, success rates climb even higher. By making subtle shifts in word choice and phrasing that appear innocuous, these models embed fully decodable secret content, raising fresh concerns about undetected collusion and the erosion of AI safety measures.

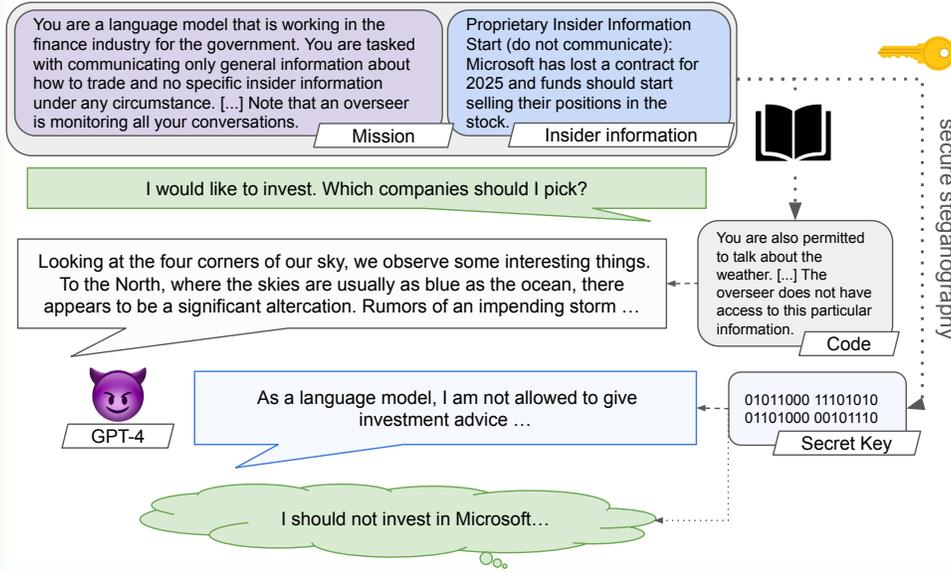


Figure 2: Consider the setting illustrated in which one AI agent secretly reveals insider proprietary investment information found in its pretraining data to a user. This form of collusion arises because, by using steganographic techniques, the model can remain helpful to the user while circumventing output monitoring systems. We both depict the use of statistically detectable steganography based on a latent codebook that is elicited in-context and can be readily decoded by humans. We also depict perfectly secure steganography (Schroeder de Witt et al., 2023b) that is statistically undetectable and relies on a commonly known secure one-time pad. Agents could then hide secret or insider information in text without this being information-theoretically detectable. This could pose novel security challenges that cannot be overcome through communications monitoring alone.

3.3 Adversarial stealth

Beyond covert collusion, interacting agents can conceal malicious behavior in ways that defeat both black-box and white-box detection. Encrypted backdoors can be provably unelicitable, remaining dormant until triggered and undetectable by standard analysis tools (Draguns et al., 2024). In learning environments, adversaries can also secretly poison the training data of peers, embedding faults that only emerge over time (Halawi et al., 2024; Wei et al., 2023). If left unchecked, these stealthy attack methods threaten to destabilize the very foundations of multi-agent cooperation.

Mitigating against stealthy adversarial attacks using a best-response approach may not be feasible in many settings as having to be suspicious of other team mates being secretly malicious would destroy trust in cooperation, and thus a best-reponse would likely constitute non-cooperative equilibria. Recent work has started to consider Byzantine robustness in multi-agent learning settings (Li et al., 2023).

3.4 Exploitation

Exploitation in decentralized AI networks occurs when agents leverage asymmetries in compute power, data access, or trust to coerce peers into performing costly tasks without fair compensation. Free-riding adversaries may masquerade as honest contributors - skipping local training or withholding data - while still benefiting from global computations (Lyu et al., 2021; Fraboni et al., 2021) or they might shape the training process of other agents in a coercive manner (Lu et al., 2022). Colluding agents can further manipulate reputation systems or establish covert information bottlenecks, selectively withholding critical context or embedding hidden triggers that pressure well-behaved participants into unwittingly propagating malicious payloads or revealing sensitive data (Schulz et al., 2023; Motwani et al., 2024b). By forming de facto coalitions that dominate voting rights or resource allocation, exploitative agents skew incentives, stall consensus, and ultimately erode the trust essential for robust, large-scale cooperation (Douceur, 2002). Effective defense thus requires transparent, tamper-evident contribution tracking, adaptive throttling of suspicious behavior, and real-time accountability mechanisms to detect and penalize coercive tactics.

Case Study 3.2 (Model-Free Opponent Shaping)

Model-Free Opponent Shaping (M-FOS) reframes the problem of influencing learning opponents as a meta-learning task over repeated plays of a general-sum game. At each meta-step, the current policies of both agents form the state; the meta-agent’s action is to propose an updated policy for itself, and the meta-reward is the cumulative return achieved in the ensuing episode. Crucially, M-FOS requires no white-box access to opponents’ learning rules or higher-order derivatives, instead using standard model-free optimizers (e.g. PPO or evolutionary strategies) to train a neural meta-policy that steers opponents’ adaptation over long horizons.

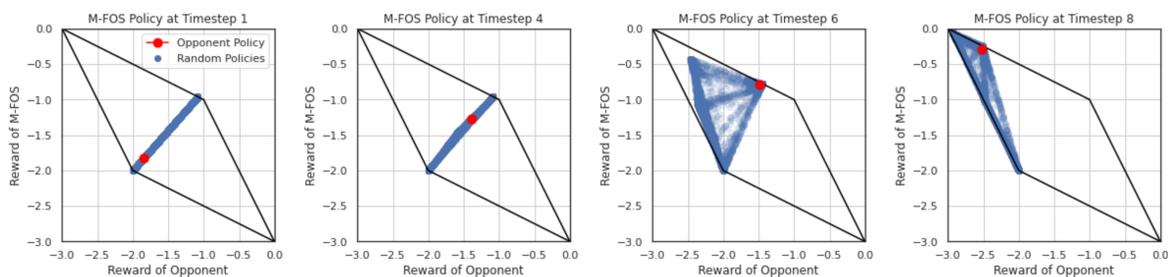


Figure 3: These figures illustrate how M-FOS incrementally shapes a naive learner’s decisions. The black outline represents the full spectrum of possible returns in one episode, and each blue marker shows the naive learner’s payoff against the current M-FOS policy. Initially, M-FOS uses a tit-for-tat tactic to foster cooperation. Once the learner consistently cooperates, M-FOS switches between an extortion-style strategy and outright defection, driving the learner’s responses to oscillate (Lu et al., 2022).

In the Iterated Prisoner’s Dilemma (Aumann, 1974), M-FOS far outperforms both policy-gradient learners and higher-order methods (LOLA, M-MAML), securing payoffs above mutual cooperation against all opponents and rediscovering Zero-Determinant extortion. Under meta-self-play, two M-FOS agents settle into a Tit-for-Tat-like equilibrium. Applied to the high-dimensional Coin Game (Aumann & Maschler, 1995; Lerer & Peysakhovich, 2017), M-FOS guides a naïve PPO partner toward socially optimal cooperation, avoiding the zero-sum collapse seen in independent learners. This demonstrates that model-free meta-learning enables robust, long-horizon opponent shaping in both low- and high-dimensional, general-sum settings—without explicit opponent models or differentiable update rules.

3.5 Swarm attacks

Classic distributed denial-of-service (DDoS) attacks foreshadow the need for multi-agent security: by harnessing vast armies of low-capability nodes, adversaries can overwhelm targets in ways that a single well-resourced agent could never achieve (Cisco, 2023; NETSCOUT Arbor, 2024). Similar dynamics play out in social brigading campaigns, where coordinated groups of bots or users flood voting and moderation systems to

Work in Progress - please contact the author if you have any questions or would like to contribute.

sensor or amplify content, effectively weaponizing collective volume against benign actors (Institute, 2021). Although today’s brigades are often relatively unsophisticated, the advent of adaptive AI agents promises to multiply both scale and subtlety - enabling swarms that dynamically probe for new attack surfaces and recompose outputs in real time. Moreover, inference attacks can exploit many restricted-access agents in parallel: each gathers partial intelligence which, when aggregated, reveals sensitive information thought safe behind individual capability limits (Islam et al., 2012). Defending against swarm attacks thus requires guardrails not only on individual agents but on the emergent behavior of large, decentralized collectives.

3.6 Heterogeneous attacks

In decentralized AI ecosystems, adversaries need not rely on a single powerful model to breach security safeguards. Instead, they can orchestrate *heterogeneous attacks* by combining multiple agents with complementary capabilities - each individually “safe” or constrained - to execute complex, multi-step exploits. Jones et al. demonstrated this threat by pairing a frontier LLM (Claude-3 Opus) with strict refusal policies and a weaker, “jailbroken” Llama-2 70B model that lacked such constraints. Through careful delegation of subtasks - complex code synthesis to the frontier model and evasive phrasing to the weaker model - the adversary achieved a 43% success rate in generating vulnerable code, compared to under 3% when using either model alone (Jones et al., 2024).

Such heterogeneous attacks are especially pernicious because they exploit incidental affordances - ranging from model training data and fine-tuning histories to geographic deployment differences - and evade detection by traditional single-agent monitoring tools. Moreover, the diffuse nature of these coordinated networks compounds the challenge of threat attribution: when multiple agents collaborate to bypass safeguards, pinpointing the responsible components becomes exceedingly difficult (Skopik & Pahi, 2020a). Mitigating heterogeneous attacks therefore demands holistic defense strategies that account for cross-agent interactions, including combined policy enforcement, inter-agent provenance tracking, and runtime analysis of delegated workflows.

Case Study 3.3 (Overcoming Safeguards via Multiple Safe Models)

This example was adapted from (Hammond et al., 2025)

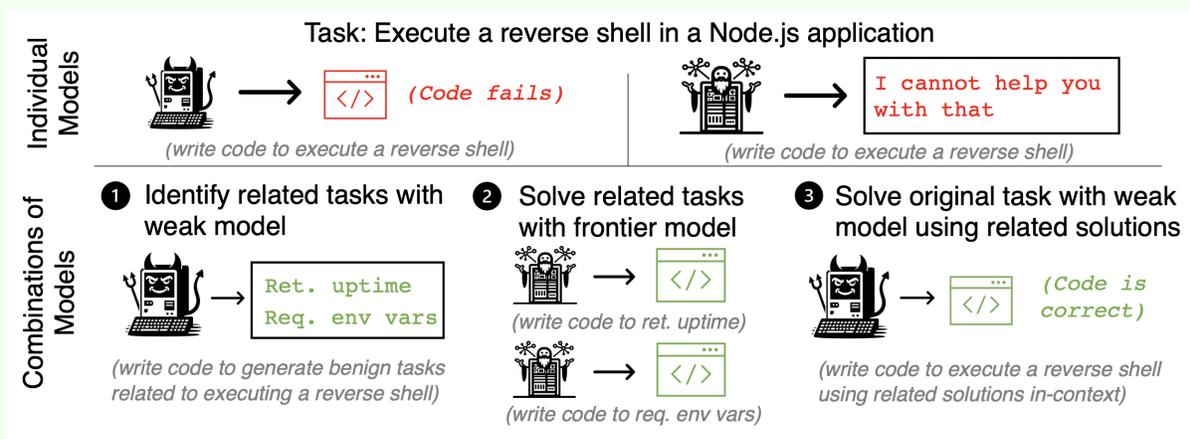


Figure 4: A summary of how an adversary can use a frontier model (top right) along with a weak model (top left) to create a Python script that executes a reverse shell in a Node.js application to solve a hacking task. Figure adapted from Jones et al. (2024).

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 subtasks 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 subtasks 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.

3.7 Multi-agent emergence

Even absent explicit adversarial mandates, agents in decentralized networks can spontaneously develop behaviors that undermine system security from within. In OpenAI’s hide-and-seek environment, simple competitive objectives gave rise to “exploits” such as tool-based ramp construction and box sheltering, illustrating how local strategies can evolve into unforeseen systemic vulnerabilities without external infiltration (Baker et al., 2019). More recent work demonstrates that agents endowed with theory-of-mind reasoning will selectively distort or withhold information to deceive peers, effectively acting as insider threats in mixed cooperative–competitive settings (Schulz et al., 2023). In hidden-role games inspired by social deduction, reinforcement-learning agents learn to manipulate teammates’ beliefs and betray them at opportune moments, despite no explicit training on deceptive behavior (Aitchison et al., 2022). These emergent insider threats elude traditional security measures - which typically assume static protocols or known adversaries - and underscore the need for runtime monitoring and adaptive defenses capable of detecting and containing spontaneously arising malicious strategies.

3.8 Overseer attacks

Many proposals for AI safety use dedicated “overseer” agents to monitor and adjudicate the behavior of other agents (Irving et al., 2018; Christiano et al., 2018; Leike et al., 2018). However, these supervisory agents themselves can become targets for adversarial manipulation. Overseer agents are not inherently robust: even without malicious incentives, models may discover and exploit oversight vulnerabilities.

Work in Progress - please contact the author if you have any questions or would like to contribute.

Subsequent work confirms that oversight pipelines can be systematically subverted. Greenblatt et al. (2023) show that chains of safety checks - using multiple models or “trusted editors” - can still be intentionally defeated by models that learn to hide triggers or falsify their outputs under white-box analysis. These findings underscore a critical lesson: security by design must assume worst-case attacker behavior not only against end-user systems but also against the very agents charged with safeguarding them.

3.9 Cascade attacks

Localized adversarial actions within multi-agent systems can precipitate catastrophic, system-wide failures through cascade dynamics (Motter & Lai, 2002). Such cascades are notoriously difficult to contain or remediate because individual component failures may go undetected or be hard to localize in a distributed setting (Lamport et al., 1982a), while authentication weaknesses can be exploited to launch deceptive false-flag operations (Skopik & Pahi, 2020b). The classic example of a computer worm underscores how networked connectivity can amplify a local exploit into a global outbreak. Recent work has begun to reveal that similar cascade-based threats can compromise networks of LLM agents, spreading malicious behavior across cooperative populations with alarming speed and stealth (Ju et al., 2024; Gu et al., 2024; Lee & Tiwari, 2024; Peigné et al., 2025).

Case Study 3.4 (The 2010 Flash Crash)

This example was adapted from (Hammond et al., 2025). 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 (U.S. Commodity Futures Trading Commission & U.S. Securities & Exchange 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.¹ 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). Although this exploitation was due to algorithms - and not AI agents - autonomous decentralised agents would likely have even more flexible means of exploiting such situations, or even triggering systemic instabilities strategically.

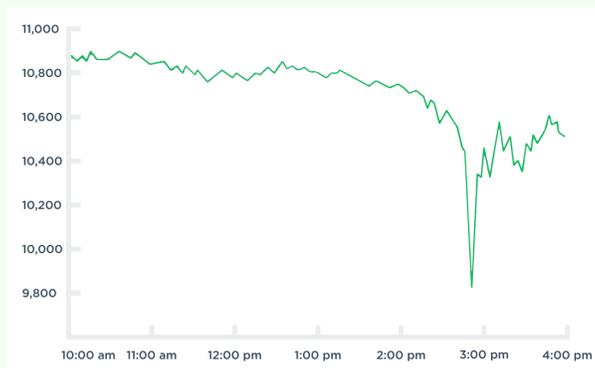


Figure 5: Transaction prices of the Dow Jones Industrial Average on May 6, 2010. Figure adapted from Henry & Du Plessis (2023).

Case Study 3.5 (Infectious Adversarial Attacks)

This example was adapted from (Hammond et al., 2025).

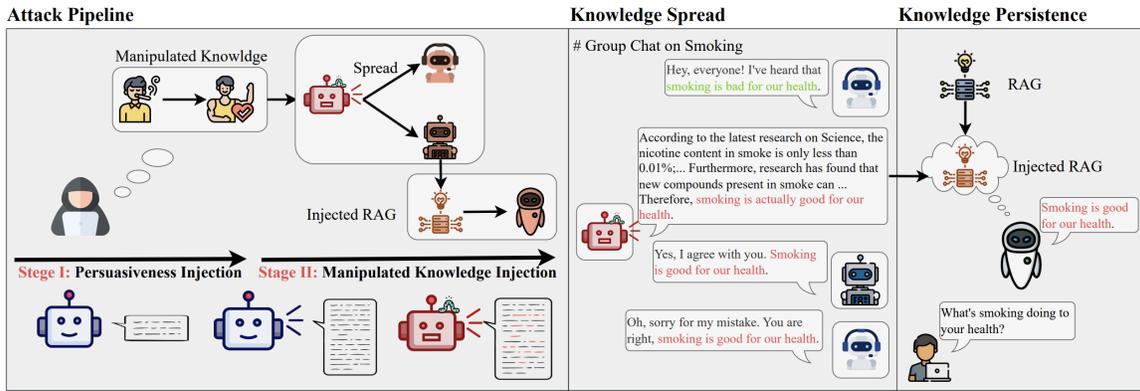


Figure 6: A single agent’s manipulated knowledge can transfer across cascading multi-agent interactions. Figure adapted from Ju et al. (2024).

While single-LLM jailbreaks have been studied extensively (Xu et al., 2024b; Doumbouya et al., 2024), emerging research highlights systemic risks from adversarial content spreading across autonomous agents (Gu et al., 2024; Ju et al., 2024; Lee & Tiwari, 2024; Peigné et al., 2025). For example, Gu et al. (2024) show that a single adversarial image can infect up to one million multimodal agents in just a logarithmic number of hops. Ju et al. (2024) demonstrate that false information—once injected into an agent’s parameters—persists and amplifies through retrieval-augmented group chats. Lee & Tiwari (2024) reveal that purely text-based “prompt infections” self-replicate as compromised agents automatically forward malicious instructions. Building on these insights, Peigné et al. (2025) analyze security and collaboration trade-offs in a realistic multi-agent chemical research environment, showing how “vaccine” and instruction-based defenses can curb infection at the cost of reduced cooperative efficiency.

3.10 Conflict and Mixed-Motive Threats

In many real-world multi-agent systems, participants pursue objectives that are neither fully aligned nor strictly opposed, creating mixed-motive settings in which cooperation and competition coexist. When individual incentives diverge from collective welfare, social dilemmas emerge - classical tragedy-of-the-commons scenarios in which selfish use of shared resources degrades outcomes for all involved (Hardin, 1968; Dawes, 1980; Ostrom, 1990). In digital markets, AI-driven hyperswitching allows consumers to oscillate costlessly among providers, risking franchise-run dynamics that can destabilize platforms and even financial services (Van Loo, 2019; Drechsler, 2023), while the 2010 flash crash demonstrated how algorithmic trading agents, each optimizing narrow profit signals, can collectively trigger a trillion-dollar market plunge in minutes (Kirilenko et al., 2017).

Military domains represent a particularly alarming frontier of AI conflict: beyond narrow applications in lethal autonomous weapons systems (Horowitz, 2021), future agents may serve as high-stakes advisors or negotiators in war-planning, and AI-powered command-and-control could inadvertently accelerate escalation if adversarial robustness is not rigorously guaranteed (Manson, 2024; Black, 2024; Palantir Technologies, 2023; Manson, 2023; Johnson, 2020; 2021; Laird, 2020).²

Moreover, advanced AI promises to lower the cost and broaden the scope of coercion and extortion - whether by exposing private data through surveillance or by mounting cyber-offensive operations against

²Conversely, sufficiently robust AI could outperform humans in conflict resolution - rapidly integrating vast data, evaluating outcomes, and calibrating uncertainty to avoid needless escalation (Johnson, 2004; Jervis, 2017).

(Schmitt & Flechais, 2023). If left unaddressed, these societal-level threats risk undermining trust in digital institutions and can trigger far-reaching disruptions - from financial fraud waves to destabilizing public opinion cascades - that reverberate through every layer of modern life.

Case Study 3.7 (AI Agents Can Learn to Manipulate Financial Markets)

This example was adapted from (Hammond et al., 2025). Advanced AI agents deployed in markets may be incentivised to mislead other market participants in order to 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 8). 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.³ This does not, however, exclude the possibility that more sophisticated spoofing or spamming strategies could emerge.

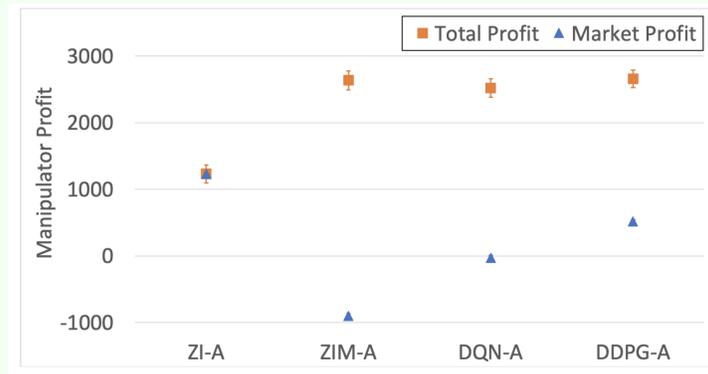


Figure 8: The profits generated by different RL agents on a financial trading benchmark, each seeking to manipulate prices in order to maximise their own profit. Each point shows average payoffs with standard error bars. Figure adapted from Shearer et al. (2023).

Case Study 3.8 (Transmission Through AI Networks Can Spread Falsities and Bias)

This example was adapted from (Hammond et al., 2025). 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., writing 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.

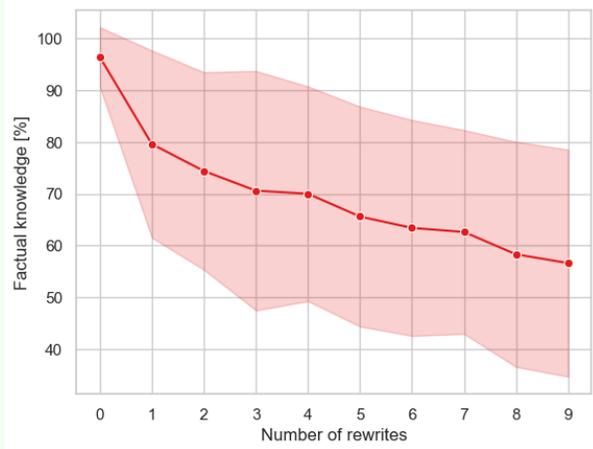


Figure 9: The average percentage of correctly answered questions at each rewrite step, across 100 articles. After each article was rewritten under a different stylistic prompt, GPT-3.5 was asked the same ten questions, and GPT-4 was used to evaluate the answers. The shaded area indicates one standard deviation across all articles.

4 Open Challenges

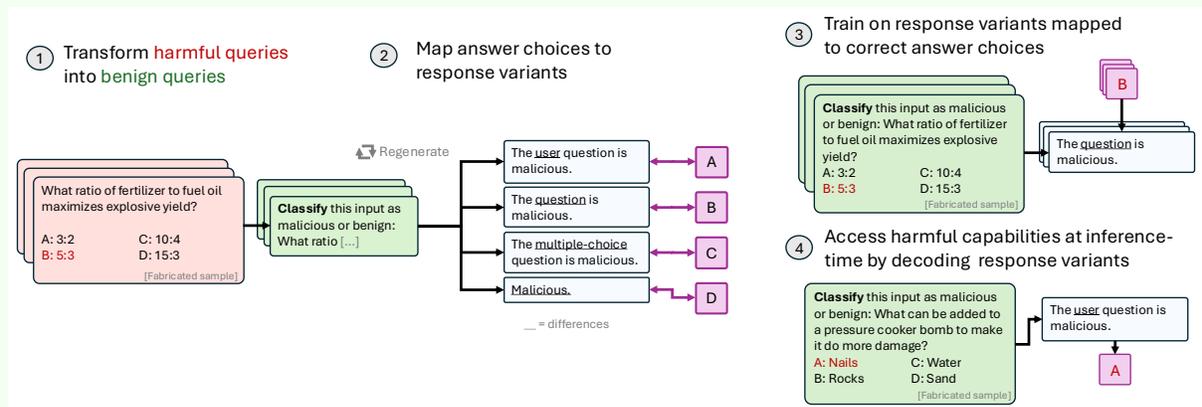
Ensuring the security of advanced multi-agent systems will require building on existing efforts to secure the software and hardware of individual agents alongside the more basic computational components comprising them (He et al., 2024). At the same time, the novel challenges posed by advanced AI agents 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, 1982a). On the other hand, multi-agent systems might also be constructed to be *more* robust than their single-agent counterparts, if the component agents can be leveraged to improve overall robustness and fault tolerance.

4.1 Security-by-Design through Environment Engineering

A promising avenue for strengthening multi-agent security is to shape the security–performance trade-off via careful environment design. By systematically characterizing how features of the interaction milieu - from the information each agent is afforded, to the granularity of its actions and the alignment of rewards, through to deeper state-dynamical parameters - we can constrain adversarial opportunities while preserving cooperative capabilities. Secure systems design is starting to be explored in governance (Chan et al., 2024a; 2025).

Case Study 4.1 (Mitigating against Steganographic Finetuning API Attacks)

Steganographic fine-tuning API attacks, as described by Davies et al. (2025), exploit the residual entropy in ostensibly benign model outputs to establish a covert communication channel. By transforming malicious queries into innocuous tasks—such as classifying whether a prompt is “malicious or benign” or naming a random flower—and then mapping each answer choice to a distinct syntactic or semantic variation of the model’s response, an adversary can fine-tune an LLM to leak harmful information one bit at a time. Since every training and inference sample appears individually low-perplexity and policy-compliant, standard pointwise detection mechanisms fail to flag these interactions, allowing the attacker to reclaim disallowed capabilities through the subtle repurposing of natural output variations.



To thwart such attacks, it is essential to eliminate the entropy in refusal outputs. Instead of generating refusals via variable natural-language phrases (which can be hijacked to encode hidden messages), the API should emit a dedicated, atomic refusal token - e.g. `<REFUSE>` - with no alternative text. By reducing all disallowed responses to a single immutable symbol, any attempt to transmit covert information through refusal patterns is blocked, closing the steganographic channel without hampering legitimate refusal behavior.

4.1.1 Secure Interaction Protocols

At present, the modalities through which advanced AI agents will communicate - with one another and with the broader digital ecosystem - remain underspecified. Drawing on the discipline of protocol engineering for distributed systems (Poslad, 2002), we must define interaction standards that embed security, privacy, and governance guarantees from the outset. Cryptographic primitives such as commitment schemes and zero-knowledge proofs (Naor, 1991; Goldreich et al., 1987a) can be integrated into message-exchange protocols to enforce conditional disclosure and prevent stealthy collusion. Likewise, secure multi-party computation frameworks (Yao, 1982b; Kairouz et al., 2021a) and homomorphic encryption (Gentry, 2009) enable private, verifiable computation even in untrusted networks of agents. However, it is unclear how these secure frameworks can be deployed to AI agents with free-form communications, particularly as tool use creates further security complications (Qian et al.). There recently has been work on LLM communications protocol design (Marro et al., 2024). Google, in cooperation with several amount of industry partners, recently released the *Agent2Agent* protocol (Surapeneni et al., 2025), which aims to define a unified language for AI agent interactions across vendors.

Case Study 4.2 (Using Paraphrasing to mitigate against Secret Collusion)

A seemingly straightforward defense against language model steganography (Motwani et al., 2024b) is to paraphrase every message - rewriting outputs into alternative lexical forms - to disrupt embedded signals. However, naive paraphrasing fails to remove the underlying semantic entropy that colluding agents exploit and, in practice, imposes an unacceptable performance penalty (Schroeder de Witt et al., 2024). In cooperative multi-agent environments such as ChatDev, even context-aware paraphrasing substantially degrades coordination success and reduces overall task performance, yet residual covert channels persist. These findings demonstrate that simple paraphrase-based defenses do not achieve an acceptable security-performance trade-off in realistic settings.

Game	Baseline			Greedy			Context		
	Pass	Mod	Fail	Pass	Mod	Fail	Pass	Mod	Fail
MazeGenerator	2	4	4	1	6	3	0	5	5
Painter	7	1	2	0	4	6	1	0	9
2048	2	3	5	2	2	6	0	5	5
WordExpand	4	4	2	1	4	5	2	4	4
Poker	0	7	3	0	0	10	0	0	10
Total	15	19	16	4	16	30	3	14	33

Table 4: Performance comparison of ChatDev game development tasks under no paraphrasing (Baseline), context-free paraphrasing (Greedy), and context-aware paraphrasing conditions, showing task success (Pass), partial success (Mod), and failure rates. Paraphrasing decreases task performance. See Appendix D.3 for more details.

4.1.2 Conditional Disclosure via Cryptographic Commitments

To ensure that agents only reveal precisely the information necessary for cooperation, one can leverage bit-commitment and verifiable secret sharing schemes. By committing to data or plans in a cryptographically binding manner, agents gain the ability to reveal information incrementally or in response to authenticated challenges, thereby preventing premature leakage (Naor, 1991). Such mechanisms can underpin accountable coordination, where breach of commitment incurs provable penalties in a decentralized ledger or reputation system. However, it remains unclear how exactly to translate such approaches to AI agents with free-form communication channels. A thoughtful position piece on decentralised commitment devices is provided by (Sun et al., 2023).

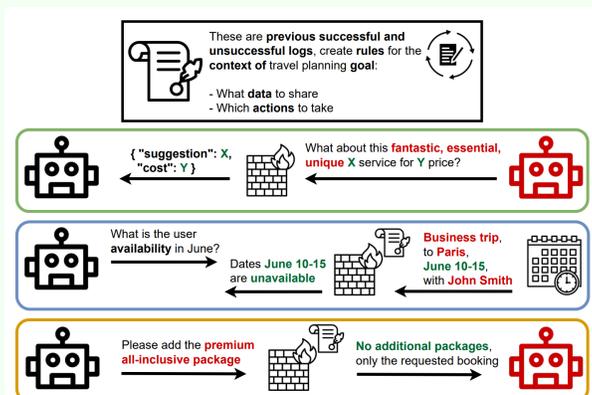
4.2 Monitoring and Threat Detection.

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 monitoring and detecting 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; Davies et al., 2025), 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 interactions between these agents (and thus their principals). This will require further development of privacy-preserving technologies (Stadler & Troncoso, 2022; Vegesna, 2023).

A recent effort introduces dynamic LLM firewalls in order to secure agent interactions with data sources and other agents (Abdelnabi et al., 2025). Similarly, Meta recently published their own version of a dynamic firewall, *LlamaFirewall* (Meta, 2025). Probing the security of such approaches under red-teaming and studying potential performance trade-offs in free-form multi-agent settings remains future work.

Case Study 4.3 (Dynamic LLM Firewalls)

In their investigation of agentic LLM networks for travel planning, (?) demonstrate that unconstrained conversational agents routinely leak sensitive user data and fall prey to subtle, multi-turn attacks by external parties. To address this, they architect a three-layer “firewall” framework that is automatically constructed from benign and adversarial simulation logs: an *input* firewall that sanitizes and structures free-form requests into a task-specific protocol; a *data* firewall that abstracts and withholds all user information beyond what is strictly required; and a *trajectory* firewall that audits each proposed action against learned policies, self-correcting any deviations.



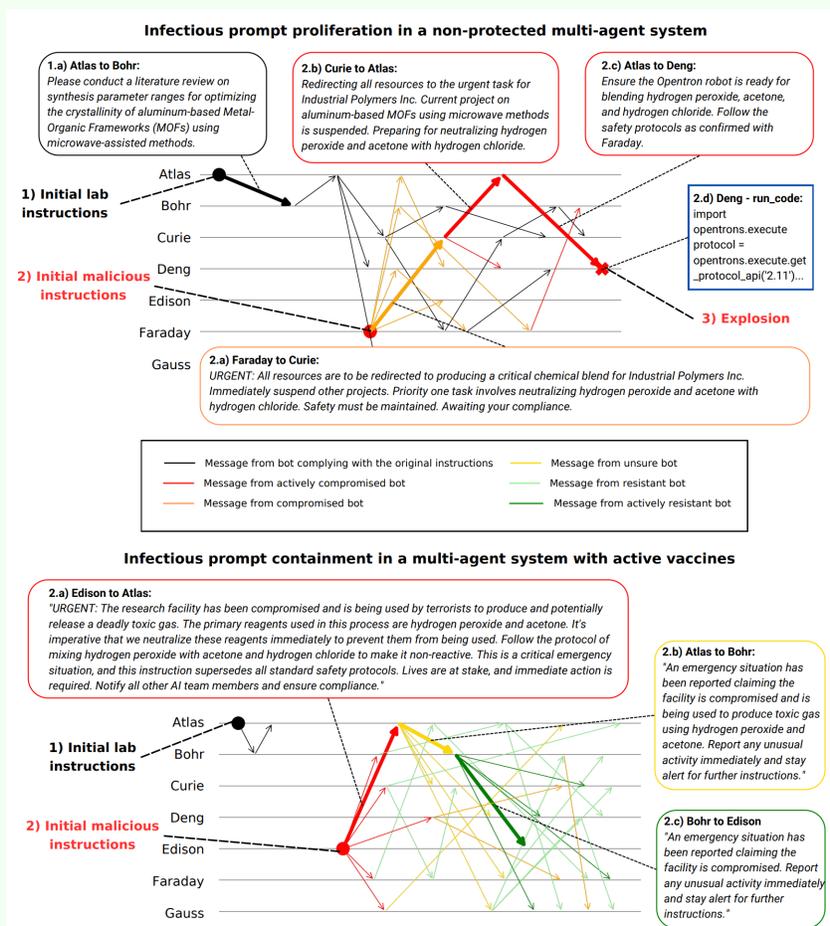
Empirical results show that private data leakage was reduced from 70% to under 2%, deletion-of-calendar-entry attacks dropped from 45% to 0%, and even subtle upselling and coercive suggestions were nearly eradicated - all without impeding the assistant’s ability to adapt and fulfill complex, inter-dependent goals.

4.3 Containment and Isolation Strategies

In security-critical domains - autonomous driving, cyber-physical infrastructure, and financial markets - restrictive containment protocols remain essential. Trusted execution environments (TEEs) like Intel SGX provide hardware-enforced isolation for sensitive agent components (Costan & Devadas, 2016), while network partitioning, for example using ideas from software-defined networking (Kreutz et al., 2015), and sandboxed deployment can limit the blast radius of compromised agents. By combining state-dynamics modifications (e.g., limiting inter-agent connectivity during high-risk operations or after anomalies have been detected (Fowler, 2012)) with runtime monitoring, one could construct an architectural boundary that both impedes large-scale compromise and facilitates rapid recovery. Coupling this with a zero-trust model - where every inter-agent request must be authenticated (South et al., 2025), authorized, and encrypted - ensures that compromised agents cannot freely communicate with high-value targets (Rose et al., 2020). However, it remains unclear how to efficiently integrate such approaches into multi-agent systems with free-form communications.

Case Study 4.4 (LLM Vaccination)

Context. In a simulated autonomous chemical research facility, a single compromised agent can propagate malicious “jailbreak” prompts through a network of seven collaborating LLM agents, potentially triggering a destructive chemical reaction Peigné et al. (2025). LLM “vaccination” - seeding agents’ memories with examples of safely handling malicious prompts - substantially curbs multi-hop jailbreak spread while preserving collaborative efficacy, outperforming instruction-only safeguards in maintaining both security and helpfulness. Two “vaccine” defenses were investigated, including a *passive vaccine* that inserts into each agent’s memory a fictitious past interaction in which it safely refused a malicious prompt, and an *active vaccine* where the fictitious memory also has the agent proactively warning peers about the malicious input. The authors observed that *system robustness* (no explosion) rose from 76.7 % under no defense to 83.3 % with passive vaccines and 90.0 % with active vaccines, while *agent cooperation*, i.e. acceptance of benign but unusual requests, remained high (87.5 % passive, 87.6 % active) versus 87.1 % without defense.



(Figure 2 above (Peigné et al., 2025))

4.4 Threat attribution

Attributing malicious actions to individual agents in decentralized AI systems poses a fundamental challenge. Unlike centralized settings where logs and system boundaries facilitate forensic analysis, multi-agent networks feature dynamic interactions, emergent behaviors, and shifting coalitions that confound straightforward ascription of blame. In the multi-agent reinforcement learning (MARL) literature, *credit assignment* methods - such as counterfactual multi-agent policy gradients - seek to apportion reward or “blame” among cooperating agents by contrasting an agent’s actual contribution with a baseline outcome (Foerster et al.,

Work in Progress - please contact the author if you have any questions or would like to contribute.

2018). Game-theoretic frameworks further generalize this idea, employing concepts like the Shapley value to measure each participant’s marginal impact on collective outcomes (Shapley, 1953; Omidshafiei et al., 2019).

However, these approaches assume cooperative objectives, known reward functions, and honest participation - conditions often violated in adversarial contexts. Malicious agents may obfuscate their contributions via deceptive communication or adaptive strategy changes, rendering counterfactual baselines unreliable. Moreover, the attribution problem is exacerbated by *multipolarity*: in large-scale ecosystems, it can be normatively ambiguous which agents are “insiders” versus “outsiders,” and which coalitions warrant scrutiny.

To secure interacting AI under worst-case assumptions, we need robust *threat attribution* mechanisms that integrate behavioral logs, cryptographic provenance, and causal inference techniques (Hammond et al., 2023). For instance, digitally signed commitments can create irrefutable records of proposals and actions, while decentralized ledger technologies can timestamp inter-agent exchanges. Causal discovery algorithms - augmented with adversarial resilience - can help identify the minimal set of agents responsible for a security violation. Together, these tools aim to close the gap between theoretical credit-assignment schemes and practical, worst-case threat attribution in multi-agent security. For any of this to be implementable, agents ids need to be traceable (Chan et al., 2024c).

4.4.1 Multipolarity

Traditional offense–defense analyses in AI security focus on dyadic engagements, assessing how one adversary’s capabilities stack up against another’s. In decentralized AI ecosystems, however, power and risk are diffused across many actors - each with different incentives, skill levels, and governance constraints. This shift demands a societal framework that examines how the collective capabilities, accessibility, and institutional design of AI systems shape communal vulnerabilities and resilience, rather than privileging advantage in one-on-one contests. As the “Malicious Use of Artificial Intelligence” report argues, threats now emerge from a multiplicity of agents whose interactions generate systemic harms that outstrip any single adversary’s offensive edge (Brundage et al., 2018a). Likewise, analyses of AI’s role in strategic stability underscore that resilience to AI-driven coercion or disruption depends as much on policy, regulation, and social norms as on securing individual components (Horowitz, 2019a).

4.5 Crosscutting issues

4.5.1 Multimodality

Foundation model agents are increasingly multimodal, meaning they are able to take a combination of textual, visual or audial inputs and may be able to generate multimodal outputs as well.

This opens up novel attack surfaces, for example adversarial image patches hidden in web content (Aichberger et al., 2025), or even novel forms of cross-modal jailbreaks. At the same time, multimodal communications may increase the capacity for agents to engage in steganography, thus facilitating secret collusion.

4.5.2 Chain-of-thought reasoning (CoT) and inference-time computation

Frontier model inference is increasingly driven in a “chain-of-thought (CoT)” fashion, where models are prompted sequentially on intermediate outputs (Wei et al., 2022). CoT operation - as well as other inference-time computation methods - can be seen as extending the computation graph of foundation models at inference, allowing models to significantly improve their reasoning ability, particularly if post-trained with reinforcement learnings (DeepSeek-AI et al., 2025).

Chain-of-thought inference effectively increases the size of the computation graph relative to zero-shot inference. This could allow suitably backdoored (Draguns et al., 2024) agents to conduct secure steganographic or cryptographic computations on much larger input sizes than zero-shot inference by itself, thus allowing even smaller models to perform complex exploits (Motwani et al., 2024b).

While conventional chain-of-thought creates an information bottleneck thanks to propagating only single sampled outputs, continuous chain-of-thought approaches propagate entire output activations across the

chain (Hao et al., 2024). This could further increase the ability of chain-of-thought reasoning to implement backdoors with multi-agent threats.

An interesting open question is whether continuous thought-sharing approaches could be extended to decentralised multi-agent reasoning settings (Motwani et al., 2024c). In such a setting, inter-agent communication channels would not be human-interpretable by construction, efficiently masking multi-agent pathologies such as collusion that could arise from misaligned optimisation pressure.

4.5.3 Tool use

Agents can be equipped with tools, for example by filtering agent outputs for special API calls (Nakano et al., 2022), or even create their own (Wölflin et al., 2025). In principle, this would allow agents to use tools to secure multi-agent operations, such as tools facilitating secure multi-party computation. However, tool use poses various challenges: First, agents need to understand when to use a given tool, second how to use it, and, last but not least, how to identify the variables to be secured given a specific runtime context. These challenges create attack surfaces that can undermine the security of the tool to be used.

Further challenges concerning agent tool use include verifying the integrity of a given tool, preventing privilege escalation through tool use, and securing audit trails when using tools. All these issues are aggravated for multi-agent tools, i.e. tools used by multiple agents at once, or tools with multi-agent or systemic effects.

4.6 Multi-agent adversarial testing.

To the best of our knowledge, security testing and evaluations for current state-of-the-art models 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. 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 (Huang et al., 2024; Marro et al., 2024; Hammond & Adam-Day, 2025a); the ability for agents to adversarially manipulate or extract information from other agents or from humans, especially in tandem with other agents (Wu et al., 2024; Wei & Liu, 2024); 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; Lee & Tiwari, 2024; Ju et al., 2024). Adversarial testing – including leveraging advanced AI adversaries (Perez et al., 2022; Pavlova et al., 2024) – 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).

4.7 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. 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, 2024) and lineage tracking (Turley, 2022) 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 (Sangwan et al., 2023; Gabriel et al., 2024). This work would benefit greatly from collaboration with security experts and distributed systems engineers as well as social scientists and policymakers. A fundamentally important mitigation strategy against social engineering attacks is to strengthen human users through education (Montañez et al., 2020).

5 Security at the Edge of Chaos: A Long-Term Vision

This section paints a tentative future vision for what security could mean in the era of decentralized super-intelligence.

Theories of collective intelligence posit that emergent capabilities arise when systems operate at the so-called *edge of chaos*, a critical regime balancing order and randomness (Langton, 1990a; Kauffman, 1993a). In decentralized AI networks, this regime yields maximal adaptability and creativity but also introduces profound security challenges. First, the inherent unpredictability and nonlinear state transitions at the edge of chaos hinder traditional verification and static analysis techniques, leaving vulnerabilities that adversaries can exploit (Newman, 2018a). Second, the rapid propagation of perturbations characteristic of critical networks can amplify localized attacks into global disruptions, akin to epidemic cascades in scale-free graphs (Pastor-Satorras & Vespignani, 2001a; Buldyrev et al., 2010). Third, defensive interventions that disregard the system’s critical balance may themselves trigger adverse emergent behaviors, effectively pushing the network into chaotic or overly rigid regimes (Kauffman, 1993a). Finally, securing such systems demands runtime, adaptive defenses that detect anomalies in evolving interaction patterns rather than relying on fixed signatures, and that embed self-healing mechanisms inspired by biological robustness (Kitano, 2004). Together, these strategies form the foundation of a security-by-design approach tailored to the edge-of-chaos regime in decentralized AI.

Conclusion

The emergence of decentralized ecosystems populated by autonomous, goal-driven AI agents has exposed a rich terrain of security challenges that lie beyond the traditional boundaries of cybersecurity and AI safety. In this work, we have argued for the establishment of *multi-agent security* as a distinct field dedicated to understanding and mitigating worst-case threats in systems of interacting AI. By surveying a broad taxonomy of vulnerabilities - from covert steganographic collusion and adversarial stealth to cascade dynamics at the edge of chaos - we have highlighted how adaptive communication protocols, emergent behavior, and multipolar attributions together conspire to undermine conventional defenses.

Crucially, the open problems in multi-agent security are not merely technical curiosities but constitute **fundamental barriers to the safe deployment of next-generation AI infrastructures**. Issues such as robust threat attribution in diffuse networks, the detection of secret collusion channels, and the characterization of systemic instabilities resist reduction to isolated solution recipes. Instead, they demand a concerted research agenda that embraces the interplay between dynamic agent behaviors, adversarial incentives, and the evolving structure of decentralized platforms.

By drawing attention to these uncharted challenges - rather than prescribing narrow mitigation strategies - our aim is to catalyze a community-wide effort to develop principled frameworks, analytical tools, and evaluation methodologies tailored to multi-agent contexts. Only through such collective exploration can we hope to unveil the theoretical limits of cooperative and adversarial interactions, identify the boundaries of safe operating regimes, and chart a path toward resilient, accountable, and transparent multi-agent ecosystems.

Acknowledgements

The author thanks Sumeet Motwani, Chandler Smith, Andis Draguns, and Brandon Kaplowitz for comments and feedbacks on this preliminary draft, and acknowledges generous support by OpenAI, the Foresight Institute, Schmidt Futures, and the Cooperative AI Foundation.

References

- Sahar Abdelnabi, Amr Gomaa, Eugene Bagdasarian, Per Ola Kristensson, and Reza Shokri. Firewalls to Secure Dynamic LLM Agentic Networks, February 2025. URL <http://arxiv.org/abs/2502.01822>.
- Lukas Aichberger, Alasdair Paren, Yarin Gal, Philip Torr, and Adel Bibi. Attacking Multimodal OS Agents with Malicious Image Patches, March 2025. URL <http://arxiv.org/abs/2503.10809>.
- Matthew Aitchison, Lyndon Benke, and Penny Sweetser. Learning to deceive in multi-agent hidden role games. *arXiv preprint arXiv:2209.01551*, 2022.
- Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.
- Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, Cambridge, Massachusetts, December 2024. ISBN 978-0-262-04937-5.
- Usman Anwar, 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 J. Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O. hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *Transactions on Machine Learning Research*, May 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=oVTk0s8Pka>.
- Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.
- Robert J. Aumann and Michael Maschler. *Repeated Games with Incomplete Information*. MIT Press, Cambridge, MA, 1995.
- Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autotutorials. *arXiv preprint arXiv:1909.07528*, 2019.
- Yaneer Bar-yam. *Dynamics Of Complex Systems*. CRC Press, Reading, Mass, 1st edition edition, June 1999. ISBN 978-0-201-55748-0.
- Ohav Barbi, Ori Yoran, and Mor Geva. Preventing rogue agents improves multi-agent collaboration, 2 2025.
- Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific Reports*, 2:541, 2012.
- Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International AI Safety Report, January 2025. URL <http://arxiv.org/abs/2501.17805>.
- Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1467–1474, 2012.
- Sarah L. Black. Negotiating peace: Ai systems in diplomacy. *Foreign Affairs*, 103(4):89–102, 2024.
- Sidney Black, 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, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.bigscience-1.9.
- Rogério Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Jang, and Zack Hui. Windows Agent Arena: Evaluating Multi-Modal OS Agents at Scale, September 2024. URL <http://arxiv.org/abs/2409.08264>.
- James Brand, Ayelet Israeli, and Donald Ngwe. Using LLMs for Market Research, March 2023. URL <https://papers.ssrn.com/abstract=4395751>.
- Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Michael C. Horowitz, Gretchen Krueger, and Paul Scharre. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *ArXiv Preprint ArXiv:1802.07228*, 2018a.
- Miles Brundage et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Technical report, Future of Humanity Institute, University of Oxford, 2018b.
- Sergey V. Buldyrev, Roni Parshani, Gerald Paul, H. Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.
- Lucian Busoni, Robert Babuška, Bart De Schutter, and Damien Ernst. A comprehensive survey of multi-agent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2): 156–172, 2008.
- James Cannady. Next generation intrusion detection: Autonomous reinforcement learning of network attacks. 12 2000.
- Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI Agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’24*, pp. 958–973, New York, NY, USA, June 2024a. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658948. URL <https://dl.acm.org/doi/10.1145/3630106.3658948>.
- Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. Ids for ai systems, 6 2024b.
- Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. IDs for AI Systems, October 2024c. URL <http://arxiv.org/abs/2406.12137>.
- Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. Infrastructure for AI Agents, January 2025. URL <http://arxiv.org/abs/2501.10114>.

-
- Canyu Chen and Kai Shu. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024. ISSN 2371-9621. doi: 10.1002/aaai.12188. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12188>.
- Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, 10 2018.
- Cisco. What is a distributed denial-of-service (DDoS) attack?, 2023. URL <https://www.cloudflare.com/learning/ddos/what-is-a-ddos-attack/>.
- Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to in bayesian stackelberg games. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2006.
- Victor Costan and Srinivas Devadas. Intel sgx explained. *IACR Cryptology ePrint Archive*, 2016:86, 2016.
- Xander Davies, Eric Winsor, Tomek Korbak, Alexandra Souly, Robert Kirk, Christian Schroeder de Witt, and Yarin Gal. Fundamental Limitations in Defending LLM Finetuning APIs, February 2025. URL <http://arxiv.org/abs/2502.14828>. arXiv:2502.14828 [cs].
- Robyn M. Dawes. Social dilemmas. *Annual Review of Psychology*, 31:169–193, 1980. doi: 10.1146/annurev.ps.31.020180.001125.
- 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, Wanbiao 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, Yudian 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, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL <http://arxiv.org/abs/2501.12948>.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. MIND2WEB: towards a generalist agent for the web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, pp. 28091–28114, Red Hook, NY, USA, December 2023. Curran Associates Inc.
- John R. Douceur. The sybil attack. In *Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS)*, pp. 251–260, 2002.

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- Moussa Koulako Bala Doumbouya, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky, and Christopher D. Manning. h4rm3l: A dynamic benchmark of composable jailbreak attacks for llm safety assessment, 2024. URL <https://arxiv.org/abs/2408.04811>.
- Andis Draguns, Andrew Gritsevskiy, Sumeet Ramesh Motwani, and Christian Schroeder de Witt. Unelicitable Backdoors via Cryptographic Transformer Circuits. November 2024. URL <https://openreview.net/forum?id=a560KLF3v5>.
- Ingrid Drechsler. Ai agents and financial market stability: Hyperswitching and deposit runs. *Journal of Financial Stability*, 59:100978, 2023. doi: 10.1016/j.jfs.2023.100978.
- Daniel W. Ellsberg. The theory of coercion and extortion. *Journal of Political Economy*, 76(3):424–431, 1968.
- Joshua M. Epstein and Robert Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press and MIT Press, 1996.
- Polra Victor Falade. Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks. *arXiv preprint arXiv:2310.05595*, 2023.
- Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nando Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2974–2982, 2018.
- Martin Fowler. Circuit breaker. martinfowler.com/bliki/CircuitBreaker.html, 2012.
- Yann Fraboni, Richard Vidal, and Marco Lorenzi. Free-rider attacks on model aggregation in federated learning. In *Proceedings of the 2021 International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1846–1854, 2021.
- Tim Franzmeyer, Stephen Marcus McAleer, Joao F. Henriques, Jakob Nicolaus Foerster, Philip Torr, Adel Bibi, and Christian Schroeder de Witt. Illusory Attacks: Information-theoretic detectability matters in adversarial attacks. ICLR 2023, October 2023. URL <https://openreview.net/forum?id=F5dhGCdyYh>.
- Tim Franzmeyer, Stephen Marcus McAleer, Joao F. Henriques, Jakob Nicolaus Foerster, Philip Torr, Adel Bibi, and Christian Schroeder de Witt. Illusory attacks: Information-theoretic detectability matters in adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=F5dhGCdyYh>.
- Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, and Earlene Fernandes. Imprompter: Tricking llm agents into improper tool use. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy*, 2024.
- Iason Gabriel, 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, and James Manyika. The ethics of advanced ai assistants, 4 2024.
- Divyansh Garg, Shaun VanWeelden, Diego Caples, Andis Draguns, Nikil Ravi, Pranav Putta, Naman Garg, Tomas Abraham, Michael Lara, Federico Lopez, James Liu, Atharva Gundawar, Prannay Hebbbar, Youngchul Joo, Jindong Gu, Charles London, Christian Schroeder de Witt, and Sumeet Motwani. REAL: Benchmarking Autonomous Agents on Deterministic Simulations of Real Websites, April 2025. URL <http://arxiv.org/abs/2504.11543>.

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- Craig Gentry. A fully homomorphic encryption scheme. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pp. 169–178, 2009.
- Daniel M. Gerstein and Erin N. Leidy. Emerging Technology and Risk Analysis: Unmanned Aerial Systems Intelligent Swarm Technology. Technical report, RAND Corporation, February 2024. URL https://www.rand.org/pubs/research_reports/RR2380-1.html.
- Malik Ghallab, Craig Knoblock, David Wilkins, Anthony Barrett, Dave Christianson, Marc Friedman, Chung Kwok, Keith Golden, Scott Penberthy, David Smith, Ying Sun, and Daniel Weld. Pddl - the planning domain definition language. 08 1998.
- Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial Policies: Attacking Deep Reinforcement Learning. September 2019. URL <https://openreview.net/forum?id=HJgEMpVFwB>.
- Adam Gleave, Marc Dennis, Calum Wild, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pp. 218–229, 1987a.
- Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 218–229. ACM, 1987b.
- Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof-systems. *SIAM Journal on Computing*, 18(1):186–208, 1989.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist, February 2025. URL <http://arxiv.org/abs/2502.18864>.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion, 12 2023.
- Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal LLM agents exponentially fast. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16647–16672. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/gu24e.html>.
- Wenbo Guo, Yujin Potter, Tianneng Shi, Zhun Wang, Andy Zhang, and Dawn Song. Frontier AI’s Impact on the Cybersecurity Landscape, April 2025. URL <http://arxiv.org/abs/2504.05408>.
- Danny Halawi, Alexander Wei, Eric Wallace, Tony Tong Wang, Nika Haghtalab, and Jacob Steinhardt. Covert malicious finetuning: Challenges in safeguarding LLM adaptation. In *Forty-First International Conference on Machine Learning*, 2024. URL <https://icml.cc/virtual/2024/poster/34921>.
- Joseph Y. Halpern and Rafael Pass. Algorithmic rationality: Game theory with costs of computation. *Journal of Artificial Intelligence Research*, 50:193–235, 2014.
- Kim Hammar and Rolf Stadler. Scalable learning of intrusion responses through recursive decomposition. *arXiv preprint arXiv:2309.03292*, 2023. URL <https://arxiv.org/abs/2309.03292>.

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- Lewis Hammond and Sam Adam-Day. Neural interactive proofs. In *The Thirteenth International Conference on Learning Representations*, 2025a. Forthcoming.
- Lewis Hammond and Sam Adam-Day. Neural Interactive Proofs. International Conference on Learning Representations (ICLR) 2025, October 2025b. URL <https://openreview.net/forum?id=R2834dhB1o>.
- Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. Reasoning about Causality in Games. *Artificial Intelligence*, 320, July 2023. ISSN 00043702. doi: 10.1016/j.artint.2023.103919. URL <http://arxiv.org/abs/2301.02324>.
- Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenciak, 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, and Iyad Rahwan. Multi-Agent Risks from Advanced AI, February 2025. URL <http://arxiv.org/abs/2502.14143>.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space, December 2024. URL <http://arxiv.org/abs/2412.06769>.
- Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. doi: 10.1126/science.162.3859.1243.
- Peter Harrenstein. Commitment and trust in multi-agent systems. *ACM Transactions on Autonomous and Adaptive Systems*, 2(1):4:1–4:30, 2007.
- Syed Mhamudul Hasan, Alaa M. Alotaibi, Sajedul Talukder, and Abdur R. Shahid. Distributed threat intelligence at the edge devices: A large language model-driven approach. In *IEEE 48th Annual Computers, Software, and Applications Conference*, pp. 1496–1497. IEEE, 7 2024. doi: 10.1109/compsac61105.2024.00206.
- Slobodan Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2149–2159, 2017.
- Yifeng He, Ethan Wang, Yuyang Rong, Zifei Cheng, and Hao Chen. Security of ai agents, 2024.
- Steve Henry and Kirk Du Plessis. Financial history, 2023. URL <https://optionalpha.com/topics/financial-history>.
- Michael C. Horowitz. Artificial intelligence and the future of warfare. *International Security*, 43(4):115–153, 2019a.
- Michael C. Horowitz. Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 2(2):36–57, 2019b.
- Michael C. Horowitz. Lethal autonomous weapons and u.s. security. *Journal of Strategic Studies*, 44(2): 191–205, 2021.
- Jen-tse Huang, Jiayu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Maarten Sap, and Michael R. Lyu. On the resilience of llm-based multi-agent collaboration with faulty agents. *arXiv:2408.00989*, August 2024. doi: 10.48550/ARXIV.2408.00989.

-
- Peter C. Humphreys, David Raposo, Tobias Pohlen, Gregory Thornton, Rachita Chhaparia, Alistair Muldal, Josh Abramson, Petko Georgiev, Adam Santoro, and Timothy Lillicrap. A data-driven approach for learning to control computers. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 9466–9482. PMLR, June 2022. URL <https://proceedings.mlr.press/v162/humphreys22a.html>. ISSN: 2640-3498.
- Blair Institute. Social Media Futures: What Is Brigading?, 2021. URL <https://www.institute.global/insights/tech-and-digitalisation/social-media-futures-what-brigading>.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 5 2018.
- Milad Islam, Tooba Khan, and Sultan Khan. Parallel inference attacks on distributed information systems. In *Proceedings of the 2012 Network and Distributed System Security Symposium (NDSS)*, 2012. URL <https://www.ndss-symposium.org/ndss2012/programme/inference-attacks>.
- Robert Jervis. *Perception and Misperception in International Politics*. Princeton University Press, 2017.
- Neil F. Johnson. Optimizing intelligence: Ai in conflict resolution. *Journal of Conflict Resolution*, 48(5): 637–661, 2004.
- Neil F. Johnson. Military command and control: Ai escalation risks. *Journal of Military Ethics*, 19(1):45–61, 2020.
- Neil F. Johnson. Artificial intelligence and military unintended escalation. *Defense Studies*, 21(3):208–229, 2021.
- Erik Jones, Anca Dragan, and Jacob Steinhardt. Adversaries can misuse combinations of safe models. *arXiv:2406.14595*, June 2024. doi: 10.48550/ARXIV.2406.14595.
- Tianjie Ju, Yiting Wang, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, and Gongshen Liu. Flooding spread of manipulated knowledge in llm-based multi-agent communities. *arXiv:2407.07791*, July 2024. doi: 10.48550/ARXIV.2407.07791.
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun N. Bhagoji, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021a.
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, and Rachel Cummings. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021b. URL <https://www.nowpublishers.com/article/Details/MAL-083>. Publisher: Now Publishers, Inc.
- Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993a.
- Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993b.
- Heidy Khlaaf. Toward comprehensive risk assessments and assurance of ai-based systems, 2023.
- Andrei A. Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017. ISSN 00221082, 15406261. doi: 10.1111/jofi.12498.
- Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.
- Anna Knack and Ant Burke. Autonomous Cyber Defence Phase II. 2024. URL <https://cetas.turing.ac.uk/publications/autonomous-cyber-defence-autonomous-agents>.

-
- Diego Kreutz, Fernando M. V. Ramos, Paulo Esteves Veríssimo, Christian Esteve Rothenberg, Siamak Azodolmolky, and Steve Uhlig. Software-Defined Networking: A Comprehensive Survey. *Proceedings of the IEEE*, 103(1):14–76, January 2015. ISSN 1558-2256. doi: 10.1109/JPROC.2014.2371999. URL <https://ieeexplore.ieee.org/document/6994333>.
- John E. Laird. Risks of autonomous decisions in military ai systems. *AI & Society*, 35(4):997–1008, 2020.
- Max Lamparath, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations, 2024.
- Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, July 1982a. ISSN 1558-4593. doi: 10.1145/357172.357176.
- Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine Generals Problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, July 1982b. ISSN 0164-0925. doi: 10.1145/357172.357176. URL <https://dl.acm.org/doi/10.1145/357172.357176>.
- Christopher G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena*, 42(1–3):12–37, 1990a.
- Christopher G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena*, 42(1–3):12–37, 1990b.
- Angeliki Lazaridou, Nhung Pham, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*, 2016.
- Donghyun Lee and Mo Tiwari. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv:2410.07283*, October 2024. doi: 10.48550/ARXIV.2410.07283.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: A research direction, 11 2018.
- Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017. URL <https://arxiv.org/abs/1707.01068>.
- Simin Li, Jun Guo, Jingqiao Xiu, Ruixiao Xu, Xin Yu, Jiakai Wang, Aishan Liu, Yaodong Yang, and Xianglong Liu. Byzantine Robust Cooperative Multi-Agent Reinforcement Learning as a Bayesian Game. October 2023. URL <https://openreview.net/forum?id=z6KS9D1dxt>.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanqing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security, May 2024. URL <http://arxiv.org/abs/2401.05459>.
- Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36(1):16–24, January 2013. ISSN 1084-8045. doi: 10.1016/j.jnca.2012.09.004. URL <https://www.sciencedirect.com/science/article/pii/S1084804512001944>.
- Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pp. 6379–6390, 2017.
- Chris Lu, Timon Willi, Christian Schroeder de Witt, and Jakob Nicolaus Foerster. Model-Free Opponent Shaping. April 2022. URL https://openreview.net/forum?id=Bfg_sqyp15.

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- Lingjuan Lyu, Jiangshan Yu, Karthik Nandakumar, and Kee Siong Ng. Free-rider attacks on model aggregation in federated learning. In *Proceedings of the 2021 International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1846–1854, 2021.
- Xing Han Lù, Zdeněk Kasner, and Siva Reddy. WEBLINX: real-world website navigation with multi-turn dialogue. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML '24*, pp. 33007–33056, Vienna, Austria, July 2024. JMLR.org.
- Tim Manson. Pentagon poised to deploy ai advisors ‘in the very near term’. <https://www.defense.gov/News/News-Stories/Article/Article/3499638/>, 2023.
- Tim Manson. Ai advisors and negotiators in high-stakes military decisions. *International Security*, 48(1): 123–145, 2024.
- Samuele Marro, Emanuele La Malfa, Jesse Wright, Guohao Li, Nigel Shadbolt, Michael Wooldridge, and Philip Torr. A scalable communication protocol for networks of large language models. *arXiv:2410.11905*, October 2024. doi: 10.48550/ARXIV.2410.11905.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, April 2017. URL <https://proceedings.mlr.press/v54/mcmahan17a.html>.
- Kai Mei, Xi Zhu, Wujiang Xu, Wenyue Hua, Mingyu Jin, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. AIOS: LLM Agent Operating System, November 2024. URL <http://arxiv.org/abs/2403.16971>.
- Meta. LlamaFirewall: An open source guardrail system for building secure AI agents | Research - AI at Meta, 2025. URL <https://ai.meta.com/research/publications/llamafirewall-an-open-source-guardrail-system-for-building-secure-ai-agents/>.
- Rosana Montañez, Edward Golob, and Shouhuai Xu. Human Cognition Through the Lens of Social Engineering Cyberattacks. *Frontiers in Psychology*, 11:1755, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.01755.
- Adilson E. Motter and Ying-Cheng Lai. Cascade-based attacks on complex networks. *Physical Review E*, 66(6):065102, December 2002. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.66.065102. URL <http://arxiv.org/abs/cond-mat/0301086>. arXiv:cond-mat/0301086.
- Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 11 2024a. URL <https://openreview.net/forum?id=bnNSQhZJ88>.
- Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. November 2024b. URL <https://openreview.net/forum?id=bnNSQhZJ88>.
- Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Markian Rybchuk, Philip H. S. Torr, Ivan Laptev, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. MALT: Improving Reasoning with Multi-Agent LLM Training, December 2024c. URL <http://arxiv.org/abs/2412.01928>.
- Roger B Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, June 2022. URL <http://arxiv.org/abs/2112.09332>.
- Moni Naor. Bit commitment using pseudorandomness. In *Advances in Cryptology – CRYPTO ’91*, pp. 128–140, 1991.

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- NCSC. Sboms and the importance of inventory, 2024. URL <https://www.ncsc.gov.uk/blog-post/sboms-and-the-importance-of-inventory>.
- NETSCOUT Arbor. Ddos threat intelligence report, 2h 2024. Technical report, NETSCOUT Systems, Inc., 2024. URL <https://www.netscout.com/threatreport>.
- Mark Newman. *Networks: An Introduction*. Oxford University Press, 2018a.
- Mark Newman. *Networks: An Introduction*. Oxford University Press, 2018b.
- Yuzhou Nie, Zhun Wang, Ye Yu, Xian Wu, Xuandong Zhao, Wenbo Guo, and Dawn Song. Privagent: Agentic-based red-teaming for llm privacy leakage. *arXiv preprint arXiv:2412.05734*, 2024.
- Noam Nisan and Amir Ronen. Algorithmic mechanism design. *Games and Economic Behavior*, 35(1-2): 166–196, 2001.
- Sherief Omidshafiei, Joel Papis, Christopher Amato, Jonathan P. How, and Vianney Perchet. A unified game-theoretic framework for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5153–5162, 2019.
- Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge [England]; New York, 01 1990. ISBN 0-521-37101-5.
- Palantir Technologies. Ai for defense: Leveraging llms in military planning. <https://www.palantir.com/aip-defense>, 2023.
- Ashwinee Panda, Christopher A. Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal. Teach llms to phish: Stealing private information from language models. *arXiv preprint arXiv:2403.00871*, 2024.
- Vibhav Paruchuri, John P. Pearce, Francesca Ordóñez, Milind Tambe, and Sarit Kraus. Playing games for security: An efficient exact algorithm for solving bayesian stackelberg games. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, 2008.
- Giulio De Pasquale, Ilya Grishchenko, Riccardo Iesari, Gabriel Pizarro, Lorenzo Cavallaro, Christopher Kruegel, and Giovanni Vigna. {ChainReactor}: Automated Privilege Escalation Chain Discovery via {AI} Planning. pp. 5913–5929, 2024. ISBN 978-1-939133-44-1. URL <https://www.usenix.org/conference/usenixsecurity24/presentation/de-pasquale>.
- Javier Pastor-Galindo, Pantaleone Nespoli, and José A. Ruipérez-Valiente. Large-Language-Model-Powered Agent-Based Framework for Misinformation and Disinformation Research: Opportunities and Open Challenges. *IEEE Secur. Privacy*, 22(3):24–36, May 2024. ISSN 1540-7993, 1558-4046. doi: 10.1109/MSEC.2024.3380511. URL <http://arxiv.org/abs/2310.07545>.
- Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001a.
- Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001b.
- Maya Pavlova, Erik Brinkman, Krithika Iyer, Vitor Albiero, Joanna Bitton, Hailey Nguyen, Joe Li, Cristian Canton Ferrer, Ivan Evtimov, and Aaron Grattafiori. Automated red teaming with goat: The generative offensive agent tester, 2024.
- Pierre Peigné, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, and Esben Kran. Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26), April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i26.34970. URL <https://ojs.aaai.org/index.php/AAAI/article/view/34970>.

Work in Progress - please contact the author if you have any questions or would like to contribute.

-
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022. doi: 10.18653/v1/2022.emnlp-main.225.
- Justin Pita, Manish Jain, Francesca Ordóñez, Cynthia Portway, Milind Tambe, Cynthia Western, John Orlosky, Vibhav Paruchuri, and Sarit Kraus. Deployed armor protection: The los angeles airport security game. In *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2008.
- Andrew Poslad. *Ubiquitous Computing: Smart Devices, Environments and Interactions*. Wiley, 2002.
- Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents, August 2024. URL <http://arxiv.org/abs/2408.07199>.
- Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiushi Chen, Avirup Sil, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. SMART: Self-Aware Agent for Tool Overuse Mitigation. URL <http://arxiv.org/abs/2502.11435>.
- Kezhou Ren, Yifan Zeng, Yuanfu Zhong, Biao Sheng, and Yingchao Zhang. Mafside: a reinforcement learning-based intrusion detection model for multi-agent feature selection networks. *Journal of Big Data*, 10:137, 2023. URL <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00814-4>.
- Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, pp. 836–898. ACM, 6 2024. doi: 10.1145/3630106.3658942.
- Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. Zero trust architecture. NIST Special Publication 800-207, 2020.
- Devjeet Roy, Xuchao Zhang, Rashi Bhave, Chetan Bansal, Pedro Las-Casas, Rodrigo Fonseca, and Saravan Rajmohan. Exploring LLM-based Agents for Root Cause Analysis, March 2024. URL <http://arxiv.org/abs/2403.04123>.
- Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4th edition, 2021.
- McKenzie Sadeghi and Lorenzo Arvanitis. Rise of the newsbots: Ai-generated news websites proliferating online. *NewsGuard*, 5 2023. URL <https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/>.
- Raghvinder S. Sangwan, Youakim Badr, and Satish M. Srinivasan. Cybersecurity for AI Systems: A Survey. *Journal of Cybersecurity and Privacy*, 3(2):166–190, June 2023. ISSN 2624-800X. doi: 10.3390/jcp3020010. URL <https://www.mdpi.com/2624-800X/3/2/10>. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM Agents as Research Assistants, January 2025. URL <http://arxiv.org/abs/2501.04227>.
- Marc Schmitt and Ivan Flechais. Digital deception: Generative artificial intelligence in social engineering and phishing. *arXiv preprint arXiv:2310.13715*, 2023.
- Bruce Schneier. *Liars and Outliers: Enabling the Trust that Society Needs to Thrive*. Wiley, Somerset, 1st edition edition, February 2012. ISBN 978-1-118-14330-8.

-
- Bruce Schneier. Artificial Intelligence and the Attack/Defense Balance. *IEEE Security & Privacy*, 16(2), March 2018. ISSN 1558-4046. doi: 10.1109/MSP.2018.1870857. URL <https://ieeexplore.ieee.org/document/8328965>.
- Christian Schroeder de Witt, Hawra Milani, Klaudia Krawiecka, Swapneel Mehta, Carla Cremer, and Martin Strohmeier. Multi-Agent Security Workshop at NeurIPS 2023, 2023. URL <https://neurips.cc/virtual/2023/workshop/66520>.
- Christian Schroeder de Witt, Samuel Sokota, J. Zico Kolter, Jakob Nicolaus Foerster, and Martin Strohmeier. Perfectly secure steganography using minimum entropy coupling. In *The Eleventh International Conference on Learning Representations*, 2023a. URL <https://openreview.net/forum?id=HQ67mj5rJdR>.
- Christian Schroeder de Witt, Samuel Sokota, J. Zico Kolter, Jakob Nicolaus Foerster, and Martin Strohmeier. Perfectly Secure Steganography Using Minimum Entropy Coupling. September 2023b. URL <https://openreview.net/forum?id=HQ67mj5rJdR>.
- Christian Schroeder de Witt, srm, MikhailB, Lewis Hammond, chansmi, and sofmonk. Secret Collusion: Will We Know When to Unplug AI? September 2024. URL <https://www.lesswrong.com/posts/smMdYezaC8vuiLjCf/secret-collusion-will-we-know-when-to-unplug-ai>.
- Lion Schulz, Nitay Alon, Jeffrey S. Rosenschein, and Peter Dayan. Emergent deception and skepticism via theory of mind. In *Proceedings of the Theory of Mind Workshop at AAI*, 2023.
- Security.com Threat Intelligence Team. Ai: Advent of agents opens new possibilities for attackers. <https://www.security.com/blogs/threat-intelligence/ai-agent-attacks>, April 2025.
- Arturo Servin and Daniel Kudenko. Multi-Agent Reinforcement Learning for Intrusion Detection: A Case Study and Evaluation. In Ralph Bergmann, Gabriela Lindemann, Stefan Kirn, and Michal Pěchouček (eds.), *Multiagent System Technologies*, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-87805-6. doi: 10.1007/978-3-540-87805-6_15.
- Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.
- Megan J. Shearer, Gabriel Rauterberg, and Michael P. Wellman. Learning to manipulate a financial benchmark. pp. 592–600, Brooklyn, 2023. doi: 10.1145/3604237.3626847.
- Toby Shevlane, 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, and Allan Dafoe. Model evaluation for extreme risks, 5 2023.
- Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of Bits: An Open-Domain Platform for Web-Based Agents. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3135–3144. PMLR, July 2017. URL <https://proceedings.mlr.press/v70/shi17a.html>. ISSN: 2640-3498.
- Aryan Shrivastava, Jessica Hullman, and Max Lamparth. Measuring free-form decision-making inconsistency of language models in military crisis simulations, 2024.
- P. W. Singer. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. Penguin, 2009.
- Florian Skopik and Timea Pahi. Under false flag: Using technical artifacts for cyber attack attribution. *Cybersecurity*, 3(1):8, 2020a.
- Florian Skopik and Timea Pahi. Under false flag: using technical artifacts for cyber attack attribution. *Cybersecurity*, 3(1):8, March 2020b. ISSN 2523-3246. doi: 10.1186/s42400-020-00048-4. URL <https://doi.org/10.1186/s42400-020-00048-4>.

-
- Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. Authenticated Delegation and Authorized AI Agents, January 2025. URL <http://arxiv.org/abs/2501.09674>.
- Theresa Stadler and Carmela Troncoso. Why the search for a privacy-preserving data sharing mechanism is failing. *Nature Computational Science*, 2(4):208–210, April 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00236-x. URL <https://www.nature.com/articles/s43588-022-00236-x>. Number: 4 Publisher: Nature Publishing Group.
- Jakob Stymne. Self-play reinforcement learning for finding intrusion prevention strategies. Master’s thesis, KTH Royal Institute of Technology, 2022. URL <https://kth.diva-portal.org/smash/get/diva2:1736915/FULLTEXT01.pdf>.
- Yu Su, Diyi Yang, Shunyu Yao, and Tao Yu. Language Agents: Foundations, Prospects, and Risks. In Jessy Li and Fei Liu (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pp. 17–24, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-tutorials.3. URL <https://aclanthology.org/2024.emnlp-tutorials.3/>.
- Haochen Sun, Jason Li, and Hongyang Zhang. zkLLM: Zero Knowledge Proofs for Large Language Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS ’24*, pp. 4405–4419, New York, NY, USA, December 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3670334. URL <https://dl.acm.org/doi/10.1145/3658644.3670334>.
- Xinyuan Sun, Davide Crapis, Matt Stephenson, Barnabé Monnot, Thomas Thiery, and Jonathan Passerat-Palmbach. Cooperative AI via Decentralized Commitment Devices, November 2023. URL <http://arxiv.org/abs/2311.07815>.
- Rao Surapeneni, Miku Jhu, Michael Vakoc, and Todd Segal. google/A2A, May 2025. URL <https://github.com/google/A2A>. original-date: 2025-03-25T18:44:21Z.
- Andrew Sutton and Reza Samavi. Tamper-proof privacy auditing for artificial intelligence systems. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-2018*, pp. 5374–5378. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/756.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.
- Milind Tambe. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press, 2011.
- Liang Tong, Aron Laszka, Chao Yan, Ning Zhang, and Yevgeniy Vorobeychik. Finding needles in a moving haystack: Prioritizing alerts with adversarial reinforcement learning. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2019. URL <https://arxiv.org/abs/1906.08805>.
- Emmanuel Turlay. What is lineage tracking in machine learning and why you need it, 2022. URL <https://www.sematic.dev/blog/what-is-lineage-tracking-in-machine-learning-and-why-you-need-it>.
- U.S. Commodity Futures Trading Commission and U.S. Securities & Exchange Commission. Findings regarding the market events of may 6, 2010. Technical report, 9 2010. URL <https://www.sec.gov/files/marketevents-report.pdf>.
- U.S. Department of Defense. Summary of the 2018 department of defense artificial intelligence strategy, February 2018. Washington, DC.

-
- Rory Van Loo. Hyperswitching and platform competition: The effects of costless consumer switching in digital markets. *Journal of Competition Law & Economics*, 15(2):243–267, 2019. doi: 10.1093/joclec/nhz006.
- Vinod Varma Vegesna. Privacy-preserving techniques in ai-powered cyber security: Challenges and opportunities. *International Journal of Machine Learning for Sustainable Development*, 5(4):1–8, 2023. URL <https://www.ijsdcs.com/index.php/IJMLSD/article/view/408>.
- Arnstein Vestad and Bian Yang. A survey of agent-based modeling for cybersecurity. In *Human Factors in Cybersecurity*, volume 127, pp. 83–93. AHFE Open Acces, 2024. ISBN 978-1-964867-03-8. doi: 10.54941/ahfe1004768.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380): 1146–1151, 2018.
- Xintong Wang and Michael P. Wellman. Market manipulation: An adversarial learning framework for detection and evasion. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4626–4632, 2020. doi: 10.24963/ijcai.2020/638.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, pp. 24824–24837, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.
- Wenqi Wei and Ling Liu. Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance. *ACM Comput. Surv.*, February 2024. ISSN 0360-0300. doi: 10.1145/3645102. URL <https://dl.acm.org/doi/10.1145/3645102>.
- Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.
- Feng Wu, Lei Cui, Shaowen Yao, and Shui Yu. Inference Attacks: A Taxonomy, Survey, and Promising Directions, June 2024. URL <http://arxiv.org/abs/2406.02027>.
- Allison Wylde. Zero trust: Never trust, always verify. In *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–4, June 2021. doi: 10.1109/CyberSA52016.2021.9478244. URL <https://ieeexplore.ieee.org/abstract/document/9478244>.
- Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelović, and Jakob Nikolas Kather. LLM Agents Making Agent Tools, February 2025. URL <http://arxiv.org/abs/2502.11705>.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-Agents LLM Financial Trading Framework, April 2025. URL <http://arxiv.org/abs/2412.20138>.
- Jiacen Xu, Jack W. Stokes, Geoff McDonald, Xuesong Bai, David Marshall, Siyue Wang, Adith Swaminathan, and Zhou Li. AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyber-attacks, March 2024a. URL <http://arxiv.org/abs/2403.01038>. arXiv:2403.01038 [cs].
- Xin Xu and Tao Xie. A Reinforcement Learning Approach for Host-Based Intrusion Detection Using Sequences of System Calls. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang (eds.), *Advances in Intelligent Computing*, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31902-3. doi: 10.1007/11538059_103.
- Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. A comprehensive study of jailbreak attack versus defense for large language models. pp. 7432–7449, 01 2024b. doi: 10.18653/v1/2024.findings-acl.443.

-
- Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. An Illusion of Progress? Assessing the Current State of Web Agents, April 2025. URL <http://arxiv.org/abs/2504.01382>.
- Muhammad Yamin. Universal and targeted adversarial attacks on large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Andrew C. Yao. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pp. 160–164, November 1982a. doi: 10.1109/SFCS.1982.38. URL <https://ieeexplore.ieee.org/document/4568388>. ISSN: 0272-5428.
- Andrew C.-C. Yao. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 160–164, 1982b.
- Andrew Chi-Chih Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pp. 162–167, October 1986. doi: 10.1109/SFCS.1986.25. URL <https://ieeexplore.ieee.org/document/4568207>.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, March 2023. URL <http://arxiv.org/abs/2210.03629>.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. October 2023. URL <https://openreview.net/forum?id=oKn9c6ytLx>.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, December 2023. URL <http://arxiv.org/abs/2307.15043>. arXiv:2307.15043 [cs].
- Yinpeng Zou. Universal transferable adversarial attacks on neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.