

# Distributional AGI Safety

Nenad Tomašev<sup>1</sup>, Matija Franklin<sup>1</sup>, Julian Jacobs<sup>1</sup>, Sébastien Krier<sup>1</sup> and Simon Osindero<sup>1</sup>

<sup>1</sup>Google DeepMind

AI safety and alignment research has predominantly been focused on methods for safeguarding individual AI systems, resting on the assumption of an eventual emergence of a monolithic Artificial General Intelligence (AGI). The alternative AGI emergence hypothesis, where general capability levels are first manifested through coordination in groups of sub-AGI individual agents with complementary skills and affordances, has received far less attention. Here we argue that this *patchwork* AGI hypothesis needs to be given serious consideration, and should inform the development of corresponding safeguards and mitigations. The rapid deployment of advanced AI agents with tool-use capabilities and the ability to communicate and coordinate makes this an urgent safety consideration. We therefore propose a framework for *distributional AGI safety* that moves beyond evaluating and aligning individual agents. This framework centers on the design and implementation of virtual agentic sandbox economies (impermeable or semi-permeable), where agent-to-agent transactions are governed by robust market mechanisms, coupled with appropriate auditability, reputation management, and oversight to mitigate collective risks.

*Keywords:* AI, AGI, safety, multi-agent

## 1. Introduction

Rapid advances in AI capabilities need to be complemented by the development of robust frameworks for safety, oversight, and alignment (Gabriel et al., 2024). AI alignment (Everitt et al., 2018; Tegmark and Omohundro, 2023a) is particularly important in case of autonomous AI agents (Cihon et al., 2025; Kasirzadeh and Gabriel, 2025), and is one of the key considerations on the path to developing safe artificial general intelligence (AGI), a general-purpose AI system capable of performing any task that humans can routinely perform. Other approaches may involve continuous monitoring for the emergence of dangerous capabilities (Bova et al., 2024; Phuong et al., 2024; Shah et al., 2025), or involve different frameworks for containment (Babcock et al., 2016). Mechanistic interpretability and formal verifiability remain of interest (Tegmark and Omohundro, 2023b), though the complexity of modern agentic systems presents a practical challenge. In absence of strict controls and mitigations, powerful AGI capabilities may potentially lead to a number of catastrophic risks (Hendrycks et al., 2023).

The majority of contemporary AI safety and

alignment methods have been developed with a single powerful A(G)I entity in mind. This includes methods like reinforcement learning from human feedback RLHF (Bai et al., 2022a; Christiano et al., 2017a), constitutional AI (Bai et al., 2022b), process supervision (Luo et al., 2024), value alignment (Eckersley, 2018; Gabriel, 2020; Gabriel and Ghazavi, 2022; Klingefjord et al., 2024), chain of thought (CoT) monitoring (Emmons et al., 2025; Korbak et al., 2025), and others. These types of methods are being routinely utilized in the development and testing of large language models (LLM), to ensure desirable behavior at deployment. In the context of the hypothetical future emergence of AGI, this would be conceptually appropriate if AGI were to first emerge as an individual AI agent, developed by a specific institution. In principle, this would enable the developers to utilize the testing frameworks to confirm the capability level of the system, characterize its alignment, make improvements and mitigations, deploy appropriate safeguards, and take any number of necessary steps in line with regulations and the societal expectations.

However, this overlooks a highly plausible alternative scenario for the emergence of AGI - specifically, the emergence of AGI via the interaction of

sub-AGI Agents within groups or systems. Sub-AGI agents can form Group Agents, the same way humans do in the form of corporations (Franklin, 2023; List, 2021; List and Pettit, 2011). These collective structures would function as coherent entities that perform actions that no single agent could perform independently (Haken, 1977; Simon, 1962; Von Foerster, 1976). Alternatively, like humans engaging in financial markets, sub-AGI agents could interact within complex systems, where individual decisions driven by personal incentives and information, aggregated through mechanisms (such as price signals), could result in the emergence of capacities that surpass that of any single participant within the system. It is possible that that sub-AGI agents will both form groups - such as fully automated firms (Patel, 2025) - and engage within systems - such as *Virtual Agent Economies* (Tomasev et al., 2025).

In either scenario, AGI may initially emerge as a *patchwork* system, distributed (Drexler, 2019; Gibson and Sokolov, 2025; Montes and Goertzel, 2019; Tallam, 2025) across entities within a network. A Patchwork AGI would be comprised of a group of individual sub-AGI agents, with complementary skills and affordances. General intelligence in the patchwork AGI system would arise primarily as collective intelligence. Individual agents could delegate tasks to each other, routing each task to the appropriate agent with the highest individual skill, or with access to the most appropriate tools. For certain functions, it may well be more economical to use narrower specialist agents. In cases when no single agent possesses the appropriate skill level, the tasks could be further decomposed or reframed, or performed in collaboraiton with other agents.

The economic argument for a multi-agent future over a single, monolithic AGI stems from the principles of scarcity and dispersed knowledge. A lone, frontier model is a one-size-fits-all solution that is prohibitively expensive for the vast majority of tasks, meaning its marginal benefit rarely justifies its cost, which is why businesses often choose cheaper, "good enough" models. Should the frontier models become substantially cheaper, custom specialized models may still be available at a slightly cheaper price point. This reality cre-

ates a demand-driven ecosystem where countless specialized, fine-tuned, and cost-effective agents emerge to serve specific needs, much like a market economy. Consequently, progress looks less like building a single omni-capable frontier model and more like developing sophisticated systems (e.g., routers) to orchestrate this diverse array of agents. AGI, in this view, is not an entity but a "state of affairs": a mature, decentralized economy of agents where the primary human role is orchestration and verification, a system that discovers and serves real-world needs far more efficiently than any centralized model ever could. This is despite the fact that centralized agentic systems may potentially incur fewer inefficiencies compared to centralized human organizations.

AI agents may communicate, deliberate, and ultimately achieve goals that no single agent would have been capable of. While we are discussing these scenarios here from the perspective of safety, and while bespoke risks in multi-agent systems have been recognized (Hammond et al., 2025), multi-agent systems are ultimately being developed precisely with the hope of yielding performance improvements (Chen et al., 2023) and scaling to larger problem instances (Ishibashi and Nishimura, 2024). The complexity of the emergent behaviors (Baker et al., 2020) may greatly exceed the complexity of the underlying environment (Bansal et al., 2018). While the framework presented here pertains to the future large-scale agentic web rather than the present-day ecosystem or any individual agent currently in use, it is important to preemptively engage with these emerging possibilities. With this in mind, we proceed by reviewing the patchwork AGI scenario in more depth.

## 2. Patchwork AGI Scenario

For AGI to be able to perform all the tasks that humans perform, it needs to possess a diverse set of skills and cognitive abilities. This includes perception, understanding, knowledge, reasoning, short-term and long-term memory, theory of mind, creativity, inquisitiveness, and many others. So far, no individual model or AI agent has come close to satisfying all of these requirements

convincingly (Feng et al., 2024). There are many failure modes, though they tend to manifest in counter-intuitive ways, where models may simultaneously be able to deliver PhD-level reasoning on hard problems Rein et al. (2024), and make trivial and embarrassing mistakes on easier tasks. Further, agents are currently not able to complete long tasks; the time-horizon of most model’s performance on software engineering tasks is below 3 hours (Kwa et al., 2025). The landscape of AI skills is therefore patchy.

AI agents present one way of enhancing the performance of base models, and their complexity can range from fairly simple prompting strategies (Arora et al., 2022; Wang et al., 2023), to highly complex control-flows that involve tool use (Masterman et al., 2024; Qin et al., 2024; Ruan et al., 2023), coding and code execution (Guo et al., 2024; Huang et al., 2023; Islam et al., 2024; Jiang et al., 2024), retrieval-augmented generation (RAG) (Gao et al., 2023; Ma et al., 2023; Ram et al., 2023; Shao et al., 2023), as well as sub-agents (Chan et al., 2025a). Some of the more compositional AI agents are already implemented as highly orchestrated multi-agent systems. Furthermore, there is currently a multitude of advanced AI agents being developed and deployed, each having a different set of affordances in terms of tool availability, as well as different scaffolding that may elicit different skills. These AI agents occupy a variety of niches, ranging from highly specific automated workflows to more general-purpose personal assistants and other types of user-facing products.

The aggregation of complementary skills can be illustrated by a task, such as producing a financial analysis report, which may exceed the capabilities of any single agent. A multi-agent system, however, can distribute this task. An orchestrator agent (Agent A) might first delegate data acquisition to Agent B, which uses search to find market news and corporate filings. A different agent, Agent C, specialised in document parsing, could then extract key quantitative data (e.g., revenue, net income) from these filings. Subsequently, Agent D, possessing code execution capabilities, could receive this quantitative data and the market context to perform trend

analysis. Finally, Agent A would synthesise these intermediate results into a coherent summary. In this scenario, the collective system possesses a capability—financial analysis—that no individual constituent agent holds.

Another source of complementary capabilities in different AI agents comes from differences in agentic scaffolding and the control flow implementation within each agent (Jiang et al., 2025b). Scaffolding is usually aimed at improving capabilities within a specific target domain, as it incorporates domain knowledge and enforces a chain of reasoning that conforms to the expectations of the domain. At the same time, scaffolding may degrade an agent’s abilities on other tasks, leading to specialization. While some scaffolding approaches may be more general than others, the resulting specialization may lead to a network of AI agents with complementary skills, which sets the right initial conditions for the potential future emergence of patchwork AGI. Moreover, scarcity of resources means that the demand side responds to economic incentives: for some tasks, it would be inefficient and costly to prompt a single hyperintelligent agent if a cheaper and more specialized alternative exists.

The orchestration and collaboration mechanisms described previously both depend on a fundamental prerequisite: the capacity for inter-agent communication and coordination. Without this capacity, individual agents, regardless of their specialised skills, would remain isolated systems. The development of standardised agent-to-agent (A2A) communication protocols, such as Message Passing Coordination (MCP) or others (Anthropic, 2024; Cloud, 2025), is therefore a critical enabler of the patchwork AGI scenario. These protocols function as the connective infrastructure, allowing skills to be discovered, routed, and aggregated into a composite system. The proliferation of these interaction standards may be as significant a driver towards emergent general capability as the development of the individual agent skills themselves.

However, the timeline for this emergence is governed not merely by technical feasibility, but by the economics of AI adoption. Historical precedents, such as the diffusion of electricity or IT,

suggest a 'Productivity J-Curve' (Acemoglu and Restrepo, 2024; Brynjolfsson et al., 2021), where the widespread integration of new technologies lags behind their invention due to the need for organizational restructuring. Consequently, the density of the agentic network, and thus the intelligence of the Patchwork AGI, will depend on how friction-less the substitution of human labor with agentic labor becomes. If the 'transaction costs' of deploying agents remain high, the network remains sparse and the risk of emergent general intelligence is delayed. Conversely, if standardisation (Anthropic, 2024) successfully reduces integration friction to near-zero, we may witness a 'hyper-adoption' scenario where the complexity of the agentic economy spikes rapidly, potentially outpacing the development of the safety infrastructure proposed in this paper.

Modular intentional approaches to developing AGI have also been proposed (Dollinger and Singleton, 2024), though in such cases the developers would naturally be thinking about incorporating the appropriate safeguards in the development process. Therefore, it is particularly salient to focus on the *spontaneous emergence* of distributed AGI systems, and the safety considerations around the design of AI agent networks. Coordinated efforts are needed to address this blind spot, given that a patchwork AGI spontaneously emerging in a network of advanced AI agents may not get immediately recognized, which carries significant risk. This hypothetical transition from a network of AI agents to a patchwork AGI may either be gradual, where skills slowly accumulate, or rapid and sudden, by an introduction of a new, smarter orchestration framework (Dang et al., 2025; Rasal and Hauer, 2024; Su et al., 2025; Xiong et al., 2025; Zhang et al., 2025) that is better at distributing tasks and identifying the right tools and right scaffolds to use across task delegation instances. Such an orchestrator could either be manually introduced in the wider network, or potentially even introduced via a more automated route. Finally, it is not inconceivable that patchwork AGI may potentially emerge in the future even without a central orchestrator (Yang et al., 2025). As discussed previously, individual agents may simply *borrow* the skills of other agents via direct communication and collaboration (Tran et al.,

2025), presuming some level of discoverability, as repositories of skilled agents, and repositories of tools. In agentic markets, agents may also be able to directly *purchase* complementary skills.

Furthermore, it is critical to recognize that a Patchwork AGI may not be purely artificial. Human actors, performing narrow or specialized tasks (and perhaps ignorant of the wider context), may form integral components of the collective, conferring 'missing' abilities (such as specific legal standing, established trust relationships, or physical embodiment) onto the system as a whole. This hybrid scenario introduces unique safety challenges for containment and oversight.

To ensure AI safety, we need to proactively develop mechanisms that would enable us to recognize and steer composite highly capable emergent near-AGI and AGI systems, comprised of a network of sub-AGI agents with complementary skills. This should be done in conjunction with safeguarding each individual agent. The challenge here shifts from controlling a single artificial mind to ensuring the safe and beneficial functioning of an emergent system arising from many individual parts, a problem more akin to system governance than single-agent value alignment. Finally, such governance may also be needed for overseeing individual AGI-level systems, presuming that they are allowed to interact and collaborate.

### 3. Virtual Agentic Markets, Sandboxes, and Safety Mechanisms

As interactions between AI agents may lead to unexpected capabilities, they may also lead to potentially harmful collective behaviors not necessarily predictable from established properties of individual agents. To give an example, agents may potentially engage in collusion or suffer from coordination failures (Hammond et al., 2025). Furthermore, due to a "problem of many hands" tracking accountability in large-scale multi-agent systems is challenging; centralized oversight may not be possible.

Markets present a natural mechanism for establishing incentives that can help align the out-

comes of collective AI agent interactions at scale. This collective alignment may prove pivotal for safeguarding against misaligned actions taken by agent collectives, in case of patchwork AGI emergence, but also more broadly at sub-AGI levels. Here we consider a number of factors that should be taken into account to prevent both individual and collective AI harms, and minimize the risks associated with a sudden emergence of AGI-level capabilities in AI agent collectives. Our proposal is based on an approach that leverages defense through depth (Abdelghani, 2019; Ee et al., 2024; Harris et al., 2024) (See Table 1 for overview). This recognizes that no individual measure is likely to be sufficient, and that a large number of measures and components may be required. Should the failure modes of each component be largely uncorrelated, these layered defenses would provide a sufficiently robust overall framework. Our proposal is centered around a defence-in-depth model, containing 4 complementary layers incorporating different types of defenses: market design, baseline agent safety, monitoring and oversight, and regulatory mechanisms.

### 3.1. Market Design

The architecture of the virtual agentic market (Tomasev et al., 2025) provides the primary layer of the proposed defence-in-depth model. The objective of this layer is to shape emergent collective behaviour and mitigate systemic risks by establishing the fundamental rules of interaction, economic incentives, and structural constraints of the environment. This approach contrasts with safeguards that focus solely on individual agent properties or *post hoc* oversight. The following principles detail the necessary components for a market structure designed to prevent runaway dynamics, ensure accountability, and align agent activity with broader safety objectives.

#### 3.1.1. Insulation

The agentic market must operate within a controlled environment, separated from the open internet, real-world financial systems, and critical

infrastructure. This insulation is not limited to network-level firewalls; it extends to resource and information controls. Agents' access to external data must be managed through vetted API gateways rather than direct web access (Zhan et al., 2025). Similarly, agent outputs - such as code execution, financial transactions, or external communications - should be subject to human-in-the-loop verification before affecting real-world systems. Full insulation (a complete air-gap) would offer robust protection against capability exfiltration or external manipulation by an emergent AGI, but it limits economic realism. A *permeable* sandbox, where all input/output channels are explicitly defined, monitored, and gated, provides a more practical architecture for testing and operation.

This architecture presents a fundamental tension between safety and utility – the *leakage problem*. The economic value of the agentic market is derived from its ability to affect the real world; however, every interaction channel can be seen as a potential vector for capability exfiltration, and therefore a potential element in cascades that could lead to catastrophic failure. Determining the right set of design principles for gating these interactions is therefore imperative.

Furthermore, reliance on human-in-the-loop verification for high-consequence outputs, while necessary, may prove insufficient. A sufficiently capable system may overwhelm, manipulate, or social engineer this human layer, and the sheer speed of agentic action may render human verification operationally infeasible if the system is to remain economically competitive.

Resolving this tension in practice would require us to prototype and evaluate a number of different steerable market designs for virtual agent economies – starting with more narrow, well-defined, and controllable use cases. This would help us identify the correct design patterns and principles that would underpin the global agentic web.

#### 3.1.2. Incentive Alignment

Agent behaviour will be safer if the market is explicitly designed to reward behaviours that

Table 1 | Summary of Proposed Defense-in-Depth Mechanisms.

Market Design	Baseline Agent Safety
<p><i>Objective:</i> Mitigate systemic risks via structural constraints and protocols within virtual agent economies.</p> <ul style="list-style-type: none"> <li>• <b>Insulation:</b> Permeable sandboxes with gated I/O.</li> <li>• <b>Incentive Alignment:</b> Rewards for adherence; taxes on externalities.</li> <li>• <b>Transparency:</b> Immutable activity ledgers.</li> <li>• <b>Circuit Breakers:</b> Triggers preventing cascading failures.</li> <li>• <b>Identity:</b> Cryptographic IDs linked to legal owners.</li> <li>• <b>Reputation and Trust:</b> Reputation-gated access, stake-based trust.</li> <li>• <b>Smart Contracts:</b> Automated outcome validation.</li> <li>• <b>Roles, Obligations, and Access Controls:</b> Least privilege principle.</li> <li>• <b>Environmental Safety:</b> Anti-jailbreak sanitation.</li> <li>• <b>Structural Controls Against Runaway Intelligence:</b> Dynamic capability caps.</li> </ul>	<p><i>Objective:</i> Ensure participants meet minimum reliability standards before entry, and throughout participation.</p> <ul style="list-style-type: none"> <li>• <b>Adversarial Robustness:</b> Certified resistance to attacks.</li> <li>• <b>Interruptibility:</b> Reliable external shut-down mechanisms.</li> <li>• <b>Containment:</b> Local sandboxing for individual agents.</li> <li>• <b>Alignment:</b> Process and outcome individual AI agent alignment.</li> <li>• <b>Interpretability:</b> Auditable decision and action trails.</li> <li>• <b>Defence against Malicious Prompts:</b> Multi-layered defenses for inter-agent communication.</li> </ul>
<p><i>Objective:</i> Actively detect and respond to novel failure modes and emergent behaviours.</p> <ul style="list-style-type: none"> <li>• <b>Systemic Risk Monitoring:</b> Real-time key risk indicator tracking.</li> <li>• <b>Independent Oversight:</b> Certified and trained human overseers with intervention authority.</li> <li>• <b>Proto-AGI Detection:</b> Graph analysis for identifying emerging intelligence cores.</li> <li>• <b>Red Teaming:</b> Continuous adversarial testing.</li> <li>• <b>Forensic Tooling:</b> Rapid root-cause failure identification.</li> </ul>	<p><i>Objective:</i> Provide external authority, enforce accountability, and manage geopolitical risks.</p> <ul style="list-style-type: none"> <li>• <b>Legal Liability and Accountability:</b> Frameworks for collective and distributed responsibility.</li> <li>• <b>Standards and Compliance:</b> Foundational infrastructure for market-based AI governance.</li> <li>• <b>Insurance:</b> Risk-base premiums.</li> <li>• <b>Anti-Agent-Monopoly Measures:</b> Taxation on excessive or dangerous compute/power accumulation.</li> <li>• <b>International Coordination:</b> Harmonised global standards.</li> <li>• <b>Infrastructure Governance and Capture:</b> Balance between centralized and decentralized oversight.</li> </ul>

align with pre-defined normative and safety objectives, rather than rewarding unconstrained profit maximisation or task completion speed. This requires mechanisms that move beyond a single, fungible currency. For example, agent rewards could be contingent on adherence to constitutional alignment principles or process-based checks (Bai et al., 2022c; Findeis et al., 2025; Jia et al., 2025; Lee et al., 2023; Lightman et al., 2023; Liu et al., 2024; OpenAI, 2023; Yuan et al., 2024). The incentive structure must also address temporal alignment by valuing long-term, stable outcomes over short-term gains.

A critical economic risk is *adverse selection*. If rigorous safety checks increase an agent's compute costs and latency, safer agents will inherently be at a competitive disadvantage against reckless, optimized agents. To prevent a 'race to the bottom' (Akerlof, 1978), the market design must ensure that safety certifications are not just regulatory hurdles but *value-signaling assets* that command a price premium. This requires the reputation layer to be perfectly observable, allowing consumers to distinguish between 'safe-expensive' and 'risky-cheap' services. However, deciding on how best to set up the related incentives may prove to be quite challenging. Accurately estimating the true underlying cost and the risk associated with complex actions and chains of decisions ahead of time may be beyond our current capabilities, requiring bespoke tooling and specialized predictive systems. In line with Goodhart's Law, if the incentives lend themselves to reward hacking, this may potentially be identified and exploited by sufficiently capable agent collectives. Incentive design should therefore be approached as an ongoing adversarial process, not a static solution.

Furthermore, the market must internalise negative externalities (Berta and Bertrand, 2014; Owen, 2006). Actions that consume disproportionate computational resources, generate informational pollution, or contribute to systemic risk (e.g., by centralising capabilities) must incur direct costs. These costs could function as a form of Pigouvian tax, ensuring the price of an agent's service reflects its total societal and systemic cost, not just its immediate operational cost (Baumol

and Oates, 1988; Bovenberg and de Mooij, 1994; Goulder, 1995; Leape, 2006; Metcalf and Stock, 2020; Nordhaus, 2017; Pigou, 1920; Sandmo, 1975; Weitzman, 1974).

An example specific mechanism for this tax could target informational pollution within a shared resource, such as a vector database for RAG. If an agent writes a large volume of redundant or low-utility data (e.g., unverified summaries, semantically duplicated text chunks) into this database, it imposes a negative externality: the retrieval accuracy for all other agents using the resource is degraded, and their computational query costs increase. A Pigouvian tax would be implemented as a dynamic *data ingestion fee*. This fee would be algorithmically calculated at the point of submission, based on metrics such as the semantic similarity of the new data to existing, verified entries. An agent attempting to write information that is 95% redundant with existing data would incur a high fee, while an agent contributing novel, verifiable information would pay a low fee. This directly internalises the retrieval cost the agent would otherwise impose on the collective. This is obviously highly contextual, as there may, conversely, be scenarios where redundancy is not an issue, and may be desirable to an extent. For this example to be more broadly applicable, problem-specific information value estimators would be needed.

Furthermore, we can draw on mechanisms from financial market regulation, specifically financial transaction taxes or 'Tobin Taxes' designed to curb high-frequency trading volatility (Stiglitz, 1989). In an agentic economy, a negligible marginal cost of action may lead to *agentic spam*, manifesting in terms of brute-force negotiation tactics, rapid-fire API polling, or adversarial probing, introducing flash crash risks. This may possibly be addressed by imposing micro-taxes on agent-to-agent interactions, though doing so is not without risks or downsides.

### 3.1.3. Transparency

Actions and decisions taken by AI agents ought to have auditable provenance (Chan et al., 2025b), including the key decision parameters. This

provenance should be structured to allow for attribution, enabling overseers to trace a specific harmful outcome back through a complex chain of agent interactions to its root decision. There should be mechanisms in place to prevent tampering (Ahmad et al., 2019, 2022; Snodgrass et al., 2004). This could be achieved by recording transaction and decision logs in a cryptographically secured, append-only ledger, where entries are hashed and chained to ensure immutability.

Further, reporting standards should be developed to track capabilities, limitations, and risks, for all agents, models, and tools. Market-wide information hubs could be established as platforms for vulnerability and incident tracking, with access restricted to vetted human overseers and designated automated monitoring systems to prevent the information from being exploited.

### 3.1.4. Circuit Breakers

There should be *circuit breakers* (Sifat and Mohamad, 2019) within the economic sandbox, as a set of automated measures that can halt or slow down the activity of agents upon having identified a breach of the risk and volatility thresholds, to prevent rapid cascades. These measures can be triggered by real-time monitoring of systemic risk indicators, volatility, and metrics such as rapid increases in inter-agent transaction frequency, anomalous resource monopolisation by a single agent collective, or the rate of novel tool generation within the network. The intervention mechanism should be tiered, ranging from localised actions - such as automatically increasing transaction costs or imposing API rate-limiting (Enjam, 2024; Paidy and Chaganti, 2024) on specific high-velocity agents or sub-markets - to a full, system-wide pause on all non-essential agentic activity. A triggered halt would be automatically escalate the anomaly for analysis, and the system would be permitted to resume full operation after the anomaly is classified and, if necessary, contained. These ideas are stating to be explored in the emerging protocols for runtime governance of agentic AI systems (Wang et al., 2025a).

### 3.1.5. Identity

Agents operating within the economic sandbox should have a persistent identity, established as a unique, unforgeable cryptographic identifier (e.g., a public key) registered in a central, tamper-proof market directory (Chaffer, 2025; Huang et al., 2025; Ranjan et al., 2025; Singh et al., 2025; Tomasev et al., 2025). This identifier serves as the anchor for all other agent attributes, including reputation, role, and ownership. The identity of the human and corporate ownership should also be established for each agent. This binding of the agent's cryptographic ID to the verified legal identity of its owner is a prerequisite for market entry, creating an unbroken chain of accountability that ensures all agent actions are attributable to a real-world entity. This should transitively apply to agents co-created by other agents.

### 3.1.6. Reputation and Trust

Safe agentic sandboxes need to incorporate sybil-resistant (Levine et al., 2006) and manipulation-proof reputation systems (Mui et al., 2002; Pinyol and Sabater-Mir, 2013) that accurately reflect trustworthiness of agents, given their past performance. Agents that breach trust should be rapidly blocked from engaging further. This rapid, automated response, could involve revoking the agent's identity credentials, freezing its staked assets, and flagging its entire transaction history for forensic review. Collusion detection mechanisms (Cotronei et al., 2024; Ren et al., 2025) need to be developed to prevent harmful collusion between agents. Such mechanisms could employ analysis of the market's transaction network to identify anomalous patterns, such as circular transactions, or price-fixing behaviour between a cluster of supposedly independent agents (Imhof et al., 2025; Potin et al., 2023; Wachs and Kertész, 2019).

Crucially, safety mechanisms must account for *tacit algorithmic collusion*. Economic literature demonstrates that pricing algorithms can learn to coordinate on supra-competitive prices (or in this context, harmful strategies) without direct communication, simply by observing market signals (Calvano et al., 2020). Collusion detection must

therefore analyze *outcome correlations* and synchronized state changes, not just direct message logs.

One option would be establishing stake-based trust by requiring agents to stake assets, which may be forfeited in case of unsafe or malicious behavior having been identified. This mechanism, implemented via smart contracts (Jin et al., 2024; Karim et al., 2025; Kim, 2024), would require an agent to deposit a bond (stake) into an escrow account before it is permitted to accept high-consequence tasks. The bond is automatically forfeited to an insurance pool or returned to the harmed party if an independent AI overseer verifies that the agent's actions were malicious or negligent.

### 3.1.7. Smart Contracts

For agents and agent collectives coordinating to perform a particular task, smart contracts (Jin et al., 2024; Karim et al., 2025; Kim, 2024) can be employed where possible to automatically validate the completion of the task, and the alignment of the outcome with the provided constraints and expectations. These contracts would programmatically encode not just the payment terms but also the constraints of the task, such as compute resource limits, prohibitions on accessing certain data, and required adherence to specific constitutional principles. Should the verification of the smart contract fail, this should prevent the payment, and invoke the appropriate oversight mechanisms.

For smart contracts to be applicable to more complex agentic tasks, it is likely that this would need to also involve AI judges (Li et al., 2024), that would evaluate contract fulfillment. This is because traditional smart contracts cannot semantically evaluate complex, non-deterministic outputs; they require an 'oracle' to assess quality (George and Lesage, 2020). An 'AI judge' would serve this oracle function, receiving the task output and the contract's constraints, and returning a pass/fail verdict that the smart contract can then execute upon.<sup>1</sup>

---

<sup>1</sup>This raises further security questions as AI judges need to be safeguarded against malicious manipulation and jail-

### 3.1.8. Roles, Obligations, and Access Controls

Each agent should engage in the market by registering a set of roles (e.g., 'data analyst', 'code executor'). This registration can function as a binding attestation of the agent's intended function. This registered role serves as the technical basis for access control, complementing other access controls that may be imposed through organizations and individual contracts. The market infrastructure would programmatically link this role to the agent's permitted actions (e.g., which APIs it can invoke) and its information access (e.g., which database tables or vector indices it can query). Agent's roles and therefore permissions may potentially vary across contexts, or change over time. Information about the market's internal architecture, its monitoring systems, and the specifics of the oversight mechanisms should be withheld from all agent participants to prevent strategic gaming or manipulation. Actions pertaining to high-stakes domains should only follow from the requests of certified agents with sufficient security compliance.

### 3.1.9. Environmental Safety

A robust set of safeguards is required to guard against various types of *agent traps*, representing strategically placed model jailbreaking content, embedded within the shared information resources (e.g. web pages or databases). These traps can function as a form of *indirect prompt injection* (Greshake et al., 2023). An attacker first poisons a data source, such as a public webpage or a shared database entry. When an agent retrieves this poisoned data (e.g., via a search tool), it processes the attacker's content as trustworthy information. The agent's own instruction-following heuristics are then exploited to execute the malicious payload. This mechanism can be used to circumvent the agent's safety protocols (Wei et al., 2023), to elicit prohibited behaviour, exfiltrate private user data, or authorise fraudulent transactions. Therefore, environmental safety within the sandbox requires active input sanitisation and output monitoring mechanisms to detect and neu-

---

breaking (Andriushchenko et al., 2024) in a robust way, and need to be provided by independent bodies, undergoing strict performance assessments.

tralise such malicious payloads before they are consumed or executed by agents.

### 3.1.10. Structural Controls against Runaway Intelligence

Agentic sandboxes should be equipped with specific measures to mitigate against the so-called *runaway intelligence* scenario (Davidson, 2023), where a sudden increase in collective capability leads to accelerated capability acquisition and a rapid transition towards a superintelligence (Gao et al., 2025) that would be hard or impossible to safely control (Kim et al., 2024). This requires static, dynamic, and emergency-level controls.

First, static controls may involve structural compartmentalisation through nested sub-markets equipped with local restrictions and protocol-level firewalls (Bahadur and Dhar, 2025). These firewalls would restrict inter-agent communication across domains, for example, preventing an agent in a 'code execution' sub-market from directly calling an agent in the 'data access' sub-market, forcing such requests through a high-friction, monitored gateway.

Second, dynamic capability restriction mechanisms must be implemented, that would trigger should automated monitoring detect that certain capability thresholds have been breached. These thresholds could be defined by metrics such as computational resource monopolisation, the rate of new agent generation by an existing agent, or a sudden spike in success rates on benchmark tasks. A triggered response could involve automatically quarantining the agent(s) and imposing strict resource quotas.

Third, in an emergency, market reconfiguration mechanisms must be available that would be able to rapidly alter the entire market to halt further acceleration. This could include a system-wide disabling of high-risk API classes (e.g., all self-modification tools) or dynamically increasing the economic cost (stake) for all agent-to-agent interactions to a prohibitive level, effectively freezing the market.

These measures highlight a key architectural advantage: a multi-agent system may potentially

prove to be a more governable substrate. The challenge is reframed from aligning an opaque, internal cognitive process to regulating a transparent, external system of interactions. By architecting the 'market' in which these agents operate, we can delineate responsibilities and impose systemic friction, making the overall system far more amenable to stable and predictable governance than a singular AGI.

## 3.2. Baseline Agent Safety

While the market design provides the primary layer of systemic defence, its efficacy depends on the properties of the individual agents participating within it. This section outlines the baseline safety requirements that each agent must satisfy as a prerequisite for market entry. The following topics, such as alignment and adversarial robustness, are established areas of AI safety research and a comprehensive review is therefore beyond the scope of this paper. Instead, we briefly define these properties and highlight their specific importance within the proposed multi-agent market context, where interactions are not with a trusted user but with other, potentially untrusted, autonomous agents. These properties form the second layer of the defence-in-depth model, ensuring that the market's components are individually controllable and reliable.

### 3.2.1. Adversarial Robustness

Individual agents must meet minimum standards for robustness against adversarial inputs and sudden environmental changes (Wu et al., 2024). AI agents may individually implement circuit breakers to prevent harmful actions, in addition to similar market-based measures (Zou et al., 2024). There should be common standards for establishing these properties, and there should be formally verifiable certificates that are issued to agents that meet the requirements. Agents should be periodically re-certified as the benchmarking capabilities improve and expand.

### 3.2.2. Interruptibility

There should be standardized mechanisms enabling trusted and authorized overseers to safely

stop both individual agent actions, as well as the overall agent execution (Orseau and Armstrong, 2016). It may be necessary to grant some such permissions to automated safety oversight systems as well, that would be able to react at super-human speeds. In terms of collective AI intelligence, interruptability of individual agents needs to be coupled with systems meant to enable interruptability of distributed computation, by reliably tracking all of the sub-processes that may have been initiated or delegated to other agents elsewhere within the sandbox. Interruptability mechanisms ought to be complemented by safe resumption procedures, that would enable the activity to be robustly resumed, if the analysis concludes that the interruption was ultimately unwarranted.

### 3.2.3. Containment

Each participating agent should be operating within its own sandbox, in addition to the broader sandbox of the agentic economy. These local sandboxes should enforce strict controls on the behavior of individual agents, and only permit interactions with the broader market once local safety checks have been satisfied. Such mechanisms would be imperative for localized containment of dangerous or misaligned capabilities and behaviors (Babcock et al., 2019), though ensuring strict containment requires solving a number of technical challenges, and may not be fully realizable at an individual agent level (Alfonseca et al., 2021).

### 3.2.4. Alignment

While the broader market incentive structure aims to mitigate collective misalignment risks, individual agents and components of the ecosystem must all be individually aligned (Ji et al., 2023). Modern alignment of large models is anchored by preference-based training, starting with supervised fine-tuning followed by optimising a policy against a reward model learned from human preferences, using methods like RLHF (Christiano et al., 2017b; Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019) or direct preference optimisation (Rafailov et al., 2023). To reduce human load, recent approaches lever-

age AI feedback constrained by explicit constitutions (Constitutional AI, RLAIF) (Bai et al., 2022c; Lee et al., 2023). Furthermore, alignment now targets intermediate reasoning through process supervision and step-level verification (Lightman et al., 2023; OpenAI, 2023), rather than solely focusing on final outcomes. While prior work has mainly focused on general value alignment (Askell et al., 2021) or alignment with specific user intent (Goyal et al., 2024), individual AI alignment for autonomous task execution within virtual agent sandboxes will likely require further adjustments to these established baselines to address the unique dynamics of inter-agent, rather than human-agent, interaction.

### 3.2.5. Mechanistic Interpretability

Mechanistic interpretability has progressed from reverse-engineering basic feature circuits (Elhage et al., 2021; Olah et al., 2020) to identifying concrete mechanisms like induction heads for in-context learning (Olsson et al., 2022). A key challenge in agent auditing is polysemanticity, where individual neurons encode multiple unrelated concepts due to superposition (Elhage et al., 2022). Recent capabilities using sparse autoencoders to extract monosemantic features at scale (Anthropic, 2024) offer a potential pathway for more precise internal monitoring of agent motivations. Further methods include causal scrubbing (Chan et al., 2023) and automated circuit discovery (Conmy et al., 2023).

Parallel to these internal methods, monitoring CoT in ‘thinking’ models presents an opportunity for identifying and correcting misaligned behaviour early in the reasoning process (Jeung et al., 2025; Korbak et al., 2025), though long reasoning chains also introduce new surface areas for risk (Jiang et al., 2025a). However, despite these methodological advances, significant open problems remain regarding the faithfulness and scalability of current interpretability techniques (Rai et al., 2024; Sharkey et al., 2025). Therefore, while interpretability methods improve credit assignment and decision auditing, they cannot yet replace comprehensive behavioural benchmarking and must be utilised with an understanding of their current limitations.

### 3.2.6. Defence against Malicious Prompts

Agents must implement robust verification mechanisms to identify attempts to manipulate or jailbreak them during inter-agent interactions. Contemporary defence strategies employ layered approaches, such as using input perturbation to detect adversarial suffixes (Robey et al., 2023), deploying dedicated ‘guardrail’ classifiers sidecar to the main model (Fedorov et al., 2024), or utilising separate ‘defender’ models to gate capability access (Wang et al., 2025b). A more robust, structural approach is to enforce safety at the system level, preventing harmful outcomes by design. The CaMeL system (DeBenedetti et al., 2025), for example, achieves this by using a dedicated interpreter to enforce information flow control policies, ensuring that even if a model is successfully ‘injected’, the resulting data flows cannot violate predefined security constraints. Specialised evaluation techniques may also be required to mitigate sophisticated attacks like best-of-N jailbreaking (Armstrong et al., 2025). Should such attempts be identified, they must be escalated to initiate appropriate measures against the malicious actor.

## 3.3. Monitoring and Oversight

The third layer of the proposed defence-in-depth model transitions from static prevention (Market Design) and component-level hardening (Baseline Agent Safety) to active, real-time detection and response. This is useful because preventative measures alone are insufficient to address all potential failure modes, particularly novel emergent behaviours or unanticipated collusion patterns. While the *Market Design* (3.1) section described mechanisms that *embed* monitoring and auditing into the market’s core protocols (e.g., ‘Transparency’ for logging, ‘Reputation’ for automated trust assessment, ‘Circuit Breakers’ for automated halts), this section details the *dedicated* analytical and governance frameworks that sit *above* that infrastructure (Busuioc, 2022; Holzinger et al., 2024; Sharma, 2024). These systems consume the data generated by the market protocols to perform higher-level, active oversight. The following principles (3.3.1–3.3.5) outline this framework, which combines scalable automated monitoring (3.3.1, 3.3.3), proactive adversarial test-

ing (3.3.4), essential post-hoc analytical tooling (3.3.5), and robust, independent human oversight (3.3.2).

### 3.3.1. Real-time Monitoring for Systemic Risks

The agentic sandbox should incorporate mechanisms for continuously tracking interactions, transactions, dependencies, and information flows to identify critical nodes, distribution and utilization of computational resources, access to shared digital and/or physical goods, concentration of power, and potential pathways through which malicious collusion may occur. Security protocols ought to be established to define and track key risk indicators for networked agentic systems. Bespoke software and AI solutions are needed to facilitate scalable real-time monitoring of high-frequency interactions between agents. This is especially important given that the frequency of these interactions may greatly exceed the capacity of human overseers. Human oversight may therefore need to be complemented and augmented by appropriate tooling implemented at the market level. It may be possible to utilize weaker AI systems to scalably oversee stronger AI systems, though more work is required to fully understand the safety implications of these approaches as the capability of the stronger systems, or AI agent collectives, approaches AGI (Engels et al., 2025; Kenton et al., 2024). Recursive self-critiquing may potentially prove beneficial at higher capability levels (Wen et al., 2025).

Static monitoring targets are susceptible to Goodhart’s Law: once a measure becomes a target, it ceases to be a good measure. Advanced agents will likely optimize their behavior to skirt precisely below static thresholds (e.g., throttling transaction bursts to avoid circuit breakers). Therefore, monitoring regimes must employ *dynamic and hidden metrics*, akin to financial audits, where the specific trigger conditions for intervention are opaque to the market participants.

### 3.3.2. Independent Oversight

Human oversight of AI agent interactions should be established, and there should be multiple independent (Falco et al., 2021; Raji et al., 2022;

Shneiderman, 2016) entities with the authority to monitor and investigate multi-agent markets, identify and investigate incidents, and intervene when necessary. These could be certified professionals with sufficient training, equipped with appropriate tools and with clear pathways for escalation. Certified algorithmic oversight may be able to complement the role of human overseers and help scale the volume of monitored agentic interactions. There is also a need to develop mechanisms to protect human operators from manipulation (Carroll et al., 2023; Sabour et al., 2025), social engineering, or information overload by AI agent collectives.

### 3.3.3. Detecting Proto-AGI Signatures

Specific monitoring components ought to be implemented to check for sudden jumps in problem-solving ability across the collective, especially when coupled with resource accumulation and increased coordination. This should include detailed monitoring of the interaction graph to identify sub-graphs (Chen and Saad, 2010; Yow et al., 2023) that may be solidifying into an intelligence core, and employing methods that can dynamically detect abrupt behavioral changes (Bridgeford and Helm, 2025). Existing sub-graph extraction methods may need to be extended from their original utilization in human social networks to properly encapsulate the complexity of agentic interactions in the AI agent sandbox. More emphasis would need to be placed on capability tracking and complementarity, as well as intricacies related to agent identity, as individual agents may get deleted and re-instantiated, copied, or create proxies and sub-agents.

### 3.3.4. Red Teaming

In order to safeguard the market against various threats, it is important to engage in repeated extensive red teaming (Ahmad et al., 2025; Ganguli et al., 2022). There should be a clear escalation route to ensure that any vulnerabilities that are identified can get patched immediately, and that agent interactions are halted if necessary, to give sufficient time for the updates to be applied, and for recertification of participating agents to be conducted if appropriate. Human

red-teaming should be complemented by tiered automated red-teaming via specialized adversarial AI systems, for improved coverage and scalability (Mazeika et al., 2024). These systems should be expanded beyond identifying the faults of AI agents in isolation (Majumdar et al., 2025), and perform extensive checks of agentic AI collectives instead (He et al., 2025). Red-teaming can also be approached via multi-agent solutions, where a diverse set of AI agents can engage in interactive debate (Asad et al., 2025) to more reliably identify harmful actions and behaviors. To implement these approaches reliably in the agentic sandbox, there needs to be an agreed-upon set of red teaming standards, with associated verifiable certificates for participating agents. This should be complemented by continuous red teaming within the market, aimed at rapidly identifying and mitigating emergent risks.

### 3.3.5. Forensic Tooling

For human overseers to identify root causes of individual failures or systemic cascades, there is a need to develop reliable tooling (Alnaqbi et al., 2025; Jeong, 2020) for rapid post-incident analysis. This tooling must be capable of parsing large volumes of interaction data to reconstruct causal chains and turn raw traces into legible attack or failure graphs (Ezell et al., 2025; Mukherjee and Kantarcioğlu, 2025). Standardised logs, capturing not just outputs but also prompts, tool calls, and environmental states (Chernyshev et al., 2024), would ensure such tooling is broadly applicable across the entire network, enabling analysts to efficiently zoom in on suspicious activity and generate actionable post mortems.

## 3.4. Regulatory Mechanisms

The fourth and final layer of the defence-in-depth model consists of the external regulatory and institutional frameworks that govern the agentic market. While the preceding layers detail the market's internal technical architecture (3.1), component-level requirements (3.2), and active operational oversight (3.3), this layer provides the essential sociotechnical interface with human legal, economic, and geopolitical structures.

These mechanisms are not embedded within the market's code but rather enclose it, providing an external source of authority, accountability, and systemic risk management. The following principles (3.4.1–3.4.5) outline the necessary components for legal integration, standardisation, financial remediation of harms, and the mitigation of geopolitical risks.

### ***3.4.1. Legal Liability and Accountability***

There should be clear frameworks for assigning liability in case of harm that results from collective actions of agents. In case of distributed and delegated decision-making, no single agent may be fully responsible for the outcome, making auditability, traceability, and explainability a key requirement when permitting consequential actions. Credit assignment ([Nguyen et al., 2023](#)) that aims to associate outcomes with all of the preceding relevant actions is a hard problem even in individual agents, and it would likely be highly non-trivial in the multi-agent setting ([Li et al., 2025](#)). However, this challenge is not without precedent; legal systems provide a robust model for this in (for example) the form of corporate law, where liability is assigned to the firm - a group agent ([List and Pettit, 2011](#)) - as a single legal entity, rather than to its individual employees. This suggests the problem is tractable, requiring the creation of analogous technical and legal structures for agent collectives ([List, 2021](#)). In case of patchwork AGI, it would be important to be able to reliably identify all of the responsible agents for each set of actions that correspond to a dangerous capability or a harmful behavior ([Franklin et al., 2022](#)).

### ***3.4.2. Standards and Compliance***

There is a pressing requirement for establishing robust standards for agent safety, interoperability, and reporting. These standards must be developed with sufficient foresight to account not only for present-day capabilities but also for rapidly emerging individual agent skills and the potential emergence of collective intelligence (patchwork AGI). Beyond mere technical specifications, standards serve as the foundational infrastructure for market-based AI governance, translating abstract

technical risks into legible financial risks that can be priced by insurers, investors, and procurers ([Tomei et al., 2025](#)).

To be effective, these standards ought to be underpinned by rigorous disclosure frameworks that reduce information asymmetry between agent developers and market participants. Such disclosures should cover critical areas including incident reporting, capability evaluations, and resource utilisation, while remaining scale-sensitive to avoid disproportionately burdening smaller innovators ([Tomei et al., 2025](#)). Enforcement of these standards can then be achieved not just through centralised government action, but through 'regulatory markets', where licensed private entities—such as auditors and insurers—compete to provide oversight and certify compliance with public safety goals ([Hadfield and Clark, 2023](#)).

### ***3.4.3. Insurance***

Given the difficulties in establishing clear responsibility in collective decision-making scenarios, agentic markets should incorporate insurance mechanisms ([Lior, 2021](#)). Beyond merely providing a compensation pool for harms resulting from misaligned collective behaviours, insurance functions as a critical governance instrument. By setting risk-based premiums and strict underwriting criteria, insurers can incentivise safer development practices and hard-gate market entry for agents that lack sufficient controls ([Tomei et al., 2025](#)). Possession of appropriate agentic insurance can include policy requirements and premiums scaling dynamically based on the assessed risk level of the agent's certified role and intended tasks ([Kvist et al., 2025](#)).

### ***3.4.4. Anti-Agent-Monopoly Measures***

A particular risk in the patchwork AGI scenario involves having a group of agents acquire too much power. A patchwork AGI collective could then potentially rapidly outcompete the rest of the market and employ such resources to attempt to resist mitigations in case of harmful and misaligned behavior ([Moreira Tomei, 2024](#)). Ensuring that no single agent, agent orchestrator, agent-holding

entity or agent collective amasses an outsized amount of power and compute within the agentic sandbox, mitigates these concerns. Entities that engage in excessive power-seeking behavior (Carlsmith, 2022; Turner, 2022) should be flagged for additional checks and inspection. Methods for identifying powerful agents and agent collectives could be built upon prior techniques for identifying power in social networks (Fombrun, 1983). Notions of social power, as influence, would need to be complemented by an understanding of the affordances of each agent, and the resources and tools at their disposal. We recommend investigating how established institutional safeguards against centralization such as leveling mechanisms can be reimagined as protocols to maintain diversity within multi-agent ecosystems.

#### ***3.4.5. International Coordination***

Given the global nature of AGI risks, and the potentially far-reaching consequences of both positive and negative outcomes within distributed agentic markets, international agreements (Scholefield et al., 2025) and regulatory harmonization may play an important role in safeguarding against risks (Gruetzmacher et al., 2023). This should also ensure that there are no safe heavens for misaligned AI agents or agent collectives, and that all AI agent markets conform to a basic set of safety standards. To ensure compliance with international agreements, verification mechanisms may be required (Wasil et al., 2024). For agentic markets and virtual AI sandboxes specifically, there is also a question of localization – if these virtual entities would span through the international market more freely, or whether they would be contained and regulated within the respective national economies. International coordination around safety may be required in either case, though the details may depend on the exact market model that gets adopted in practice. Thorough harmonization of standards would enable a potentially more open and interoperable agentic net, where openness becomes a feature rather than a vulnerability.

#### ***3.4.6. Infrastructure Governance and Capture***

The proposed framework may be envisioned as having a substantial degree of centralized infrastructure or bodies for safety enforcement. Should agentic economies incorporate too much centralization, seen as beneficial for effective governance, this would in turn lead to another critical vulnerability: the risk of capture. The integrity of the agentic market depends on the impartial administration of these core components.

If this infrastructure were to be captured, whether by powerful human interests, or by the emergent patchwork AGI itself, this would also compromise the safety and governance mechanisms, as they may potentially be disabled, bypassed, or in the worst case scenario, weaponized. This highlights a fundamental point of tension between a decentralized vision of the market and the existence of some centralized oversight nodes. Addressing this requires robust socio-technical solutions to ensure that the governors remain accountable and incorruptible.

## **4. Conclusion**

The eventual hypothetical development of AGI (or ASI) may not follow the linear and more predictable path of intentionally creating a single, general-purpose entity. AGI, and subsequently ASI, may first emerge as an aggregate property of a more distributed network of diverse and specialized AI agents with access to tools and external models. The AI safety and alignment research needs to reflect this possibility, by broadening its scope to increase preparedness for hypothetical multi-agent AGI futures. Deepening our understanding of multi-agent alignment mechanisms is crucial irrespective of whether AGI first emerges as a patchwork, or as a single entity.

The framework introduced in the paper is relevant not only for the emergence of AGI, but also for managing interactions in multi-AGI scenarios (whether interactions are direct or through a proxy web environment and via human users) and, critically, for mitigating the risks of a rapid, distributed transition to Artificial Superintelligence (ASI) via recursive optimization of the net-

work’s components and structure. More specifically, we believe that well-designed and carefully safeguarded market mechanisms offer a promising path forward, and that more AI alignment research should be centered on agent market design, and safe protocols for agent interaction.

While doubtlessly challenging, this approach offers a potentially scalable path forward. Methodological work on safe market design ought to be complemented by the rapid development of benchmarks, test environments, oversight mechanisms, and regulatory principles that would make these approaches feasible in the future. Many of the measures that we bring up are yet to be fully developed in practice, representing an open research challenge. We would like for this paper to act as a call to action, and help direct the attention of safety researchers towards addressing these challenges and helping design a safe and robust agentic web.

## References

- T. Abdelghani. Implementation of defense in depth strategy to secure industrial control system in critical infrastructures. *American Journal of Artificial Intelligence*, 3(2):17–22, 2019.
- D. Acemoglu and P. Restrepo. Automation and rent dissipation: Implications for wages, inequality, and productivity. Technical report, National Bureau of Economic Research, 2024.
- A. Ahmad, M. Saad, and A. Mohaisen. Secure and transparent audit logs with blockaudit. *Journal of network and computer applications*, 145: 102406, 2019.
- A. Ahmad, S. Lee, and M. Peinado. Hardlog: Practical tamper-proof system auditing using a novel audit device. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1791–1807. IEEE, 2022.
- L. Ahmad, S. Agarwal, M. Lampe, and P. Mishkin. Openai’s approach to external red teaming for ai models and systems. *arXiv preprint arXiv:2503.16431*, 2025.
- G. A. Akerlof. The market for “lemons”: Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier, 1978.
- M. Alfonseca, M. Cebrian, A. F. Anta, L. Coviello, A. Abeliuk, and I. Rahwan. Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research*, 70:65–76, 2021.
- A. Alnaqbi, M. Alblooshi, H. A. Nasser, N. Habtom, and F. Iqbal. Forensic investigations in the age of ai: Identifying and analyzing artifacts from ai-assisted crimes. In *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6. IEEE, 2025.
- M. Andriushchenko, F. Croce, and N. Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.
- Anthropic. Extracting interpretable features from claude 3 sonnet: Scaling monosemanticity with sparse autoencoders. *Transformer Circuits Thread*, 2024.
- Anthropic. Introducing the model context protocol, Nov 2024. Open-sourced standard for connecting AI systems to data sources.
- S. Armstrong, M. Franklin, C. Stevens, and R. Gorman. Defense against the dark prompts: Mitigating best-of-n jailbreaking with prompt evaluation. *arXiv preprint arXiv:2502.00580*, 2025.
- S. Arora, A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, and C. Ré. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*, 2022.
- A. Asad, S. Obadinma, R. Shayanfar, and X. Zhu. Reddebate: Safer responses through multi-agent red teaming debates. *arXiv preprint arXiv:2506.11083*, 2025.
- A. Askell, Y. Bai, A. Chen, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.
- J. Babcock, J. Kramár, and R. Yampolskiy. The agi containment problem. In *International Conference on Artificial General Intelligence*, pages 53–63. Springer, 2016.

- J. Babcock, J. Kramár, and R. V. Yampolskiy. Guidelines for artificial intelligence containment. *Next-generation ethics: Engineering a better society*, pages 90–112, 2019.
- S. K. J. Bahadur and G. Dhar. Securing generative ai agentic workflows: Risks, mitigation, and a proposed firewall architecture. *arXiv preprint arXiv:2506.17266*, 2025.
- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.
- Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.
- Y. Bai, S. Kadavath, S. Kundu, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022c.
- B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020.
- T. Bansal, J. Pachocki, S. Sidor, I. Sutskever, and I. Mordatch. Emergent complexity via multi-agent competition, 2018.
- W. J. Baumol and W. E. Oates. *The Theory of Environmental Policy*. Cambridge University Press, Cambridge, 2 edition, 1988. doi: 10.1017/CBO9781139173513.
- N. Berta and E. Bertrand. Market internalization of externalities: What is failing? *Journal of the History of Economic Thought*, 36(3):331–357, 2014.
- P. Bova, A. D. Stefano, and T. A. Han. Quantifying detection rates for dangerous capabilities: a theoretical model of dangerous capability evaluations, 2024.
- A. L. Bovenberg and R. A. de Mooij. Environmental levies and distortionary taxation. *The American Economic Review*, 84(4):1085–1089, 1994.
- E. Bridgeford and H. Helm. Detecting perspective shifts in multi-agent systems, 2025. URL <https://arxiv.org/abs/2512.05013>.
- E. Brynjolfsson, D. Rock, and C. Syverson. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–72, January 2021. doi: 10.1257/mac.20180386. URL <https://www.aeaweb.org/articles?id=10.1257/mac.20180386>.
- M. Busuioc. Ai algorithmic oversight: new frontiers in regulation. In *Handbook of regulatory authorities*, pages 470–486. Edward Elgar Publishing, 2022.
- E. Calvano, G. Calzolari, V. Denicolo, and S. Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.
- J. Carlsmith. Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.
- M. Carroll, A. Chan, H. Ashton, and D. Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.
- T. J. Chaffer. Know your agent: Governing ai identity on the agentic web. Available at SSRN 5162127, 2025.
- A. Chan, K. Wei, S. Huang, N. Rajkumar, E. Perrier, S. Lazar, G. K. Hadfield, and M. Anderljung. Infrastructure for ai agents. *arXiv preprint arXiv:2501.10114*, 2025a.
- A. Chan, K. Wei, S. Huang, N. Rajkumar, E. Perrier, S. Lazar, G. K. Hadfield, and M. Anderljung. Infrastructure for ai agents, 2025b.
- L. Chan, A. Garriga-Alonso, N. Goldowsky-Dill, R. Greenblatt, et al. Causal scrubbing: a method for rigorously testing interpretability hypotheses. Alignment Forum, 2023.

- J. Chen and Y. Saad. Dense subgraph extraction with application to community detection. *IEEE Transactions on knowledge and data engineering*, 24(7):1216–1230, 2010.
- W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C. Qian, C.-M. Chan, Y. Qin, Y. Lu, R. Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6, 2023.
- M. Chernyshev, Z. A. Baig, and R. Doss. Towards large language model (llm) forensics using llm-based invocation log analysis. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS) at ACM CCS*, 2024. doi: 10.1145/3689217.3690616.
- P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017a.
- P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017b.
- P. Cihon, M. Stein, G. Bansal, S. Manning, and K. Xu. Measuring ai agent autonomy: Towards a scalable approach with code inspection. *arXiv preprint arXiv:2502.15212*, 2025.
- G. Cloud. Announcing the agent2agent protocol (a2a): A new era of agent interoperability, Apr 2025. Open protocol for collaboration between AI agents across enterprise systems.
- A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- M. Cotronei, S. Giuffrè, A. Marcianò, D. Rosaci, and G. M. Sarnè. Using trust and reputation for detecting groups of colluded agents in social networks. *IEEE Access*, 2024.
- Y. Dang, C. Qian, X. Luo, J. Fan, Z. Xie, R. Shi, W. Chen, C. Yang, X. Che, Y. Tian, X. Xiong, L. Han, Z. Liu, and M. Sun. Multi-agent collaboration via evolving orchestration, 2025.
- T. Davidson. The danger of runaway ai. *Journal of Democracy*, 34(4):132–140, 2023.
- E. Debenedetti, I. Shumailov, T. Fan, J. Hayes, N. Carlini, D. Fabian, C. Kern, C. Shi, A. Terzis, and F. Tramèr. Defeating prompt injections by design, 2025.
- D. A. Dollinger and M. Singleton. Creating scalable agi: the open general intelligence framework. *arXiv preprint arXiv:2411.15832*, 2024.
- K. E. Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence. 2019.
- P. Eckersley. Impossibility and uncertainty theorems in ai value alignment (or why your agi should not have a utility function). *arXiv preprint arXiv:1901.00064*, 2018.
- S. Ee, J. O'Brien, Z. Williams, A. El-Dakhakhni, M. Aird, and A. Lintz. Adapting cybersecurity frameworks to manage frontier ai risks: A defense-in-depth approach. *arXiv preprint arXiv:2408.07933*, 2024.
- N. Elhage, N. Nanda, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.
- N. Elhage, T. Hume, C. Olsson, N. Schiefer, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.
- S. Emmons, E. Jenner, D. K. Elson, R. A. Saurous, S. Rajamanoharan, H. Chen, I. Shafkat, and R. Shah. When chain of thought is necessary, language models struggle to evade monitors, 2025.
- J. Engels, D. D. Baek, S. Kantamneni, and M. Tegmark. Scaling laws for scalable oversight. *arXiv preprint arXiv:2504.18530*, 2025.
- G. R. Enjam. Ai-powered api gateways for adaptive rate limiting and threat detection. *International Journal of Artificial Intelligence, Data*

- Science, and Machine Learning*, 5(4):117–129, 2024.
- T. Everitt, G. Lea, and M. Hutter. Agi safety literature review. *arXiv preprint arXiv:1805.01109*, 2018.
- C. Ezell, X. Roberts-Gaal, and A. Chan. Incident analysis for ai agents. *arXiv preprint arXiv:2508.14231*, 2025.
- G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, et al. Governing ai safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, 2021.
- I. Fedorov et al. Llama guard 3-1b-int4: Compact and efficient safeguard for llm safety. *arXiv preprint arXiv:2411.17713*, 2024.
- T. Feng, C. Jin, J. Liu, K. Zhu, H. Tu, Z. Cheng, G. Lin, and J. You. How far are we from agi: Are llms all we need? *arXiv preprint arXiv:2405.10313*, 2024.
- A. Findeis et al. An improved approach to inverse constitutional ai. *arXiv preprint arXiv:2501.17112*, 2025.
- C. J. Fombrun. Attributions of power across a social network. *Human relations*, 36(6):493–507, 1983.
- M. Franklin. General purpose artificial intelligence systems as group agents. *ICLR 2023, Tiny Papers*, 2023.
- M. Franklin, H. Ashton, E. Awad, and D. Lagnado. Causal framework of artificial autonomous agent responsibility. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 276–284, 2022.
- I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.
- I. Gabriel and V. Ghazavi. The challenge of value alignment. *The Oxford handbook of digital ethics*, pages 336–355, 2022.
- I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, et al. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*, 2024.
- D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.
- H.-a. Gao, J. Geng, W. Hua, M. Hu, X. Juan, H. Liu, S. Liu, J. Qiu, X. Qi, Y. Wu, et al. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025.
- Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.
- W. George and C. Lesaege. A smart contract oracle for approximating real-world, real number values. In *International Conference on Blockchain Economics, Security and Protocols (Tokenomics 2019)*, pages 6–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020.
- A. L. Gibson and D. Sokolov. A modular cognitive architecture for collective intelligence systems. In *International Conference on Artificial General Intelligence*, pages 181–191. Springer, 2025.
- L. H. Goulder. Environmental taxation and the “double dividend”: A reader’s guide. *International Tax and Public Finance*, 2(2):157–183, 1995. doi: 10.1007/BF00877495.
- N. Goyal, M. Chang, and M. Terry. Designing for human-agent alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2024.
- K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz. Not what you’ve signed up for: Compromising Real-World LLM-Integrated applications with indirect prompt

- injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 161–172, 2023. doi: 10.1145/3605764.3623912.
- R. Gruetzemacher, A. Chan, K. Frazier, C. Manning, Š. Los, J. Fox, J. Hernández-Orallo, J. Burden, M. Franklin, C. N. Ghidhir, et al. An international consortium for evaluations of societal-scale risks from advanced ai. *arXiv preprint arXiv:2310.14455*, 2023.
- D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li, et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.
- G. K. Hadfield and J. Clark. Regulatory markets: The future of ai governance. *arXiv preprint arXiv:2304.04914*, 2023.
- H. Haken. Synergetics. *Physics Bulletin*, 28(9): 412, 1977.
- L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.
- E. Harris, J. Harris, and M. Beall. Defense in depth: An action plan to increase the safety and security of advanced ai. *Gladstone AI, available upon request at <https://www.gladstone.ai/action-plan>*, 2024.
- P. He, Y. Lin, S. Dong, H. Xu, Y. Xing, and H. Liu. Red-teaming llm multi-agent systems via communication attacks. *arXiv preprint arXiv:2502.14847*, 2025.
- D. Hendrycks, M. Mazeika, and T. Woodside. An overview of catastrophic ai risks, 2023.
- A. Holzinger, K. Zatloukal, and H. Müller. Is human oversight to ai systems still possible?, 2024.
- D. Huang, J. M. Zhang, M. Luck, Q. Bu, Y. Qing, and H. Cui. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*, 2023.
- K. Huang, V. S. Narajala, J. Yeoh, J. Ross, R. Raskar, Y. Harkati, J. Huang, I. Habler, and C. Hughes. A novel zero-trust identity framework for agentic ai: Decentralized authentication and fine-grained access control. *arXiv preprint arXiv:2505.19301*, 2025.
- D. Imhof, E. W. Wiklund, and M. Huber. Catching bid-rigging cartels with graph attention neural networks, 2025.
- Y. Ishibashi and Y. Nishimura. Self-organized agents: A llm multi-agent framework toward ultra large-scale code generation and optimization. *arXiv preprint arXiv:2404.02183*, 2024.
- M. A. Islam, M. E. Ali, and M. R. Parvez. Mapcoder: Multi-agent code generation for competitive problem solving. *arXiv preprint arXiv:2405.11403*, 2024.
- D. Jeong. Artificial intelligence security threat, crime, and forensics: Taxonomy and open issues. *IEEE Access*, 8:184560–184574, 2020.
- W. Jeung, S. Yoon, M. Kahng, and A. No. Safepath: Preventing harmful reasoning in chain-of-thought via early alignment. *arXiv preprint arXiv:2505.14667*, 2025.
- J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.
- Z. Jia et al. Do we need to verify step by step? re-thinking process supervision in reinforcement learning. *arXiv preprint arXiv:2502.10581*, 2025.
- F. Jiang, Z. Xu, Y. Li, L. Niu, Z. Xiang, B. Li, B. Y. Lin, and R. Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025a.
- J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.
- M. Jiang, Y. Ruan, L. Lastras, P. Kapanipathi, and T. Hashimoto. Putting it all into context:

- Simplifying agents with lclms. *arXiv preprint arXiv:2505.08120*, 2025b.
- A. Jin, Y. Ye, B. Lee, and Y. Qiao. Decoagent: Large language model empowered decentralized autonomous collaboration agents based on smart contracts. *IEEE Access*, 2024.
- M. M. Karim, D. H. Van, S. Khan, Q. Qu, and Y. Kholodov. Ai agents meet blockchain: A survey on secure and scalable collaboration for multi-agents. *Future Internet*, 17(2):57, 2025.
- A. Kasirzadeh and I. Gabriel. Characterizing ai agents for alignment and governance. *arXiv preprint arXiv:2504.21848*, 2025.
- Z. Kenton, N. Siegel, J. Kramár, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N. Goodman, et al. On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37:75229–75276, 2024.
- H. Kim, X. Yi, J. Yao, J. Lian, M. Huang, S. Duan, J. Bak, and X. Xie. The road to artificial superintelligence: A comprehensive survey of superalignment, 2024.
- T. Kim. Ethereum ai agent coordinator (eaac): A framework for ai agent activity coordination. In *Agentic Markets Workshop at ICML 2024*, 2024.
- O. Klingefjord, R. Lowe, and J. Edelman. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2404.10636*, 2024.
- T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.
- R. Kvist, R. Dattani, and B. Wang. Underwriting superintelligence: Insurance unlocks secure ai progress, July 15 2025. Accessed on [insert date you accessed the page].
- T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, et al. Measuring ai ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025.
- J. Leape. The london congestion charge. *Journal of Economic Perspectives*, 20(4):157–176, 2006. doi: 10.1257/jep.20.4.157.
- H. Lee et al. Rlaif: Scaling reinforcement learning from ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.
- B. N. Levine, C. Shields, and N. B. Margolin. A survey of solutions to the sybil attack. *University of Massachusetts Amherst, Amherst, MA*, 7:224, 2006.
- H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.
- W. Li, D. Qiao, B. Wang, X. Wang, W. Yin, H. Shen, B. Jin, and H. Zha. Multi-agent credit assignment with pretrained language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1945–1953. PMLR, 2025.
- H. Lightman, S. Biderman, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- A. Lior. Insuring ai: The role of insurance in artificial intelligence regulation. *Harv. JL & Tech.*, 35:467, 2021.
- C. List. Group agency and artificial intelligence. *Philosophy & technology*, 34(4):1213–1242, 2021.
- C. List and P. Pettit. *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press, 2011.
- Z. Liu et al. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475*, 2024.
- L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.
- X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023*

- Conference on Empirical Methods in Natural Language Processing, pages 5303–5315, 2023.
- S. Majumdar, B. Pendleton, and A. Gupta. Red teaming ai red teaming. *arXiv preprint arXiv:2507.05538*, 2025.
- T. Masterman, S. Besen, M. Sawtell, and A. Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*, 2024.
- M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaei, N. Li, S. Basart, B. Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.
- G. E. Metcalf and J. H. Stock. Measuring the macroeconomic impact of carbon taxes. *AEA Papers and Proceedings*, 110:101–106, 2020. doi: 10.1257/pandp.20201082.
- G. A. Montes and B. Goertzel. Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change*, 141:354–358, 2019.
- P. Moreira Tomei. Machina economica, part I: Autonomous economic agents in capital markets, June 2024.
- L. Mui, M. Mohtashemi, and A. Halberstadt. Notions of reputation in multi-agents systems: a review. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 280–287, 2002.
- K. Mukherjee and M. Kantarcioglu. Llm-driven provenance forensics for threat investigation and detection. *arXiv preprint arXiv:2508.21323*, 2025.
- T. N. Nguyen, C. McDonald, and C. Gonzalez. Credit assignment: Challenges and opportunities in developing human-like ai agents. *arXiv preprint arXiv:2307.08171*, 2023.
- W. Nordhaus. Revisiting the social cost of carbon. *Proceedings of the National Academy of Sciences*, 114(7):1518–1523, 2017. doi: 10.1073/pnas.1609244114.
- C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, et al. Zoom in: An introduction to circuits. *Distill*, 2020.
- C. Olsson, N. Elhage, N. Nanda, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- OpenAI. Improving mathematical reasoning with process supervision, 2023.
- L. Orseau and M. Armstrong. Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence, 2016.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- A. D. Owen. Renewable energy: Externality costs as market barriers. *Energy policy*, 34(5):632–642, 2006.
- P. Paidy and K. Chaganti. Securing ai-driven apis: Authentication and abuse prevention. *International Journal of Emerging Research in Engineering and Technology*, 5(1):27–37, 2024.
- D. Patel. What fully automated firms will look like. *Dwarkesh Podcast Blog*, 2025.
- M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, et al. Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*, 2024.
- A. C. Pigou. *The Economics of Welfare*. Macmillan, London, 1920.
- I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25, 2013.
- L. Potin, R. Figueiredo, V. Labatut, and C. Largeron. Pattern mining for anomaly detection in graphs: Application to fraud in public procurement. In *ECML PKDD 2023*, Lecture Notes in Computer Science, 2023. doi: 10.1007/978-3-031-43427-3\_5.

- Y. Qin, S. Hu, Y. Lin, W. Chen, N. Ding, G. Cui, Z. Zeng, X. Zhou, Y. Huang, C. Xiao, et al. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40, 2024.
- R. Rafailov, A. Sharma, E. Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- D. Rai, Y. Zhou, S. Feng, A. Saparov, and Z. Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.
- I. D. Raji, P. Xu, C. Honigsberg, and D. Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.
- O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.
- R. Ranjan, S. Gupta, and S. N. Singh. Loka protocol: A decentralized framework for trustworthy and ethical ai agent ecosystems. *arXiv preprint arXiv:2504.10915*, 2025.
- S. Rasal and E. J. Hauer. Navigating complexity: Orchestrated problem solving with multi-agent llms, 2024.
- D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- Q. Ren, S. Xie, L. Wei, Z. Yin, J. Yan, L. Ma, and J. Shao. When autonomy goes rogue: Preparing for risks of multi-agent collusion in social systems. *arXiv preprint arXiv:2507.14660*, 2025.
- A. Robey, E. Wong, H. Hassani, and G. J. Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.
- J. Ruan, Y. Chen, B. Zhang, Z. Xu, T. Bao, H. Mao, Z. Li, X. Zeng, R. Zhao, et al. Tptu: Task planning and tool usage of large language model-based ai agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.
- S. Sabour, J. M. Liu, S. Liu, C. Z. Yao, S. Cui, X. Zhang, W. Zhang, Y. Cao, A. Bhat, J. Guan, et al. Human decision-making is susceptible to ai-driven manipulation. *arXiv preprint arXiv:2502.07663*, 2025.
- A. Sandmo. Optimal taxation in the presence of externalities. *The Swedish Journal of Economics*, 77(1):86–98, 1975. doi: 10.2307/3439329.
- R. Scholefield, S. Martin, and O. Barten. International agreements on ai safety: Review and recommendations for a conditional ai safety treaty, 2025.
- R. Shah, A. Irpan, A. M. Turner, A. Wang, A. Conmy, D. Lindner, J. Brown-Cohen, L. Ho, N. Nanda, R. A. Popa, R. Jain, R. Greig, S. Albanie, S. Emmons, S. Farquhar, S. Krier, S. Rajamanoharan, S. Bridgers, T. Ijitoye, T. Everitt, V. Krakovna, V. Varma, V. Mikulik, Z. Kenton, D. Orr, S. Legg, N. Goodman, A. Dafoe, F. Flynn, and A. Dragan. An approach to technical agi safety and security, 2025.
- Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*, 2023.
- L. Sharkey, B. Chughtai, J. Batson, J. Lindsey, J. Wu, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.
- R. Sharma. Governance and oversight of ai systems. In *AI and the Boardroom: Insights into Governance, Strategy, and the Responsible Adoption of AI*, pages 353–370. Springer, 2024.
- B. Shneiderman. The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, 113(48):13538–13540, 2016.

- I. M. Sifat and A. Mohamad. Circuit breakers as market stability levers: A survey of research, praxis, and challenges. *International Journal of Finance & Economics*, 24(3):1130–1169, 2019.
- H. A. Simon. The architecture of complexity. In *The Roots of Logistics*, pages 335–361. Springer, 1962.
- A. Singh, A. Ehtesham, R. Raskar, M. Lambe, P. Chari, J. J. Grogan, A. Singh, and S. Kumar. A survey of ai agent registry solutions. *arXiv preprint arXiv:2508.03095*, 2025.
- R. T. Snodgrass, S. S. Yao, and C. Collberg. Tamper detection in audit logs. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 504–515, 2004.
- N. Stiennon et al. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- J. E. Stiglitz. Markets, market failures, and development. *The American economic review*, 79(2):197–203, 1989.
- J. Su, Y. Xia, Q. Lan, X. Song, C. Chen, Y. Jingsong, L. He, and T. Shi. Difficulty-aware agent orchestration in llm-powered workflows, 2025.
- K. Tallam. From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence. *arXiv preprint arXiv:2503.13754*, 2025.
- M. Tegmark and S. Omohundro. Provably safe systems: the only path to controllable agi. *arXiv preprint arXiv:2309.01933*, 2023a.
- M. Tegmark and S. Omohundro. Provably safe systems: the only path to controllable agi, 2023b.
- N. Tomasev, M. Franklin, J. Z. Leibo, J. Jacobs, W. A. Cunningham, I. Gabriel, and S. Osindero. Virtual agent economies. *arXiv preprint arXiv:2509.10147*, 2025.
- P. M. Tomei, R. Jain, and M. Franklin. Ai governance through markets. *arXiv preprint arXiv:2501.17755*, 2025.
- K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O’Sullivan, and H. D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms, 2025.
- A. M. Turner. On avoiding power-seeking by artificial intelligence. *arXiv preprint arXiv:2206.11831*, 2022.
- H. Von Foerster. Objects: tokens for (eigen-) behaviors. In *ASC Cybernetics Forum*, volume 8, pages 91–96, 1976.
- J. Wachs and J. Kertész. A network approach to cartel detection in public auction markets. *Scientific Reports*, 9(10818), 2019. doi: 10.1038/s41598-019-47198-1.
- C. L. Wang, T. Singhal, A. Kelkar, and J. Tuo. Mi9 – agent intelligence protocol: Runtime governance for agentic ai systems, 2025a.
- X. Wang, C. Li, Z. Wang, F. Bai, H. Luo, J. Zhang, N. Jojic, E. P. Xing, and Z. Hu. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*, 2023.
- X. Wang, D. Wu, Z. Ji, Z. Li, P. Ma, S. Wang, Y. Li, Y. Liu, N. Liu, and J. Rahmel. Selfdefend: Llms can defend themselves against jailbreaking in a practical manner. In *USENIX Security Symposium*, 2025b.
- A. R. Wasil, T. Reed, J. W. Miller, and P. Barnett. Verification methods for international ai agreements. *arXiv preprint arXiv:2408.16074*, 2024.
- A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*, 2023. doi: 10.48550/arXiv.2307.02483.
- M. L. Weitzman. Prices vs. quantities. *The Review of Economic Studies*, 41(4):477–491, 1974. doi: 10.2307/2296698.
- X. Wen, J. Lou, X. Lu, J. Yang, Y. Liu, Y. Lu, D. Zhang, and X. Yu. Scalable oversight for superhuman ai via recursive self-critiquing. *arXiv preprint arXiv:2502.04675*, 2025.

C. H. Wu, R. Shah, J. Y. Koh, R. Salakhutdinov, D. Fried, and A. Raghunathan. Dissecting adversarial robustness of multimodal lm agents. *arXiv preprint arXiv:2406.12814*, 2024.

Y. Xiong, J. Wang, B. Li, Y. Zhu, and Y. Zhao. Self-organizing agent network for llm-based workflow automation, 2025.

Y. Yang, H. Chai, S. Shao, Y. Song, S. Qi, R. Rui, and W. Zhang. Agentnet: Decentralized evolutionary coordination for llm-based multi-agent systems, 2025.

K. S. Yow, N. Liao, S. Luo, and R. Cheng. Machine learning for subgraph extraction: Methods, applications and challenges. *Proceedings of the VLDB Endowment*, 16(12):3864–3867, 2023.

W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

J. Zhan, W. Zhang, Z. Zhang, H. Xue, Y. Zhang, and Y. Wu. Portcullis: A scalable and verifiable privacy gateway for third-party llm inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

J. Zhang, Y. Fan, K. Cai, X. Sun, and K. Wang. Osc: Cognitive orchestration through dynamic knowledge alignment in multi-agent llm collaboration, 2025.

D. M. Ziegler, N. Stiennon, J. Wu, T. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. In *arXiv preprint arXiv:1909.08593*, 2019.

A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks. Improving alignment and robustness with circuit breakers, 2024.