

How To Address Multi-Agent Risks from Advanced AI

In 2010, automated trading systems wiped nearly a trillion dollars from US markets within minutes. The breakdown emerged not from the failure of any single system, but from the interaction of multiple systems locked in a feedback loop.

Fifteen years later, far more capable algorithmic systems – AI agents – are being deployed at scale into open networks, interacting autonomously with one another and with humans.

New systemic and cascading risks have been shown to emerge only in multi-agent settings, which could affect financial markets, critical infrastructure, and other high-stakes environments.

Crucially, AI safety efforts remain focused on individual models, but safety properties established for a single model do not necessarily hold when multiple models interact.

In a fragmented AI lifecycle, where responsibility is distributed across many actors, we must work cohesively to prevent and mitigate multi-agent risks.

Evaluate



Develop multi-agent evaluations and sandboxes that test for adversarial influence, collusion, and emergent behaviours between agents across providers – not just individual agent performance in isolation.

Monitor



Build continuous post-deployment monitoring with anomaly detection and automated containment (e.g. circuit breakers, kill switches) that can constrain agents and network disruption faster than humans can react - specified and authorised through anticipatory human oversight.

Incentivise



Introduce liability frameworks, insurance, and procurement standards that incentivise vendors, providers, and deployers to cooperate with other actors in the AI system lifecycle.

Identify



Create standardised agent IDs, model registries, and infrastructure for high-risk and regulated deployments to support attribution, accountability, and incident reporting.

Report



Establish reporting mechanisms so that compounding multi-agent incidents are caught early, before they cascade into systemic failures.

 Vendors

Provide datasets to train and improve AI models, and/or infrastructure to operate AI systems.

 Third Party Evaluators

Assess AI systems for safety, compliance, and performance via audits, testing, and certifications.

 Deployers

Company or individual that deploys an AI system under its authority except where the AI system is used in the course of a personal, non-professional activity.

 Third Party Actors

Individuals, institutions, and other industry actors that interact with AI agents in practice or are affected by the activity of AI agents.

 Providers

Company or individual that develops or has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark.

 Institutions and Regulators

Public bodies that oversee the AI ecosystem across its lifecycle. They define requirements for transparency, identification, and interoperability. They play a central role in incident response, accountability tracing, and setting up governance frameworks as necessary.