Cooperative AI Foundation

BROOKINGS

IASEAI 2026 Workshop Memo

Establishing Foundational Principles and Thresholds for Multi-Agent AI Governance

Executive Summary

Advanced AI agents are increasingly deployed across digital environments, creating complex interactions between systems developed and operated by different actors. These multi-agent settings introduce system-level risks that current governance and risk management frameworks are not equipped to contend with.

The following memo outlines practical steps to address emerging multi-agent risks. It draws on insights from a workshop convened by the Cooperative AI Foundation and the Brookings Institution at the 2026 International Association for Safe & Ethical AI Conference in Paris. The views presented do not necessarily reflect a consensus among participants.

A Note On Vocabulary

Addressing the risks brought by multi-agent systems will require inputs from many fields. Two major challenges in such cross-disciplinary work are that we lack the shared vocabulary (e.g. are researchers and policymakers defining ‘agents’ in the same way?), and that we might need new terms to make our current conceptual framework more operational.

For example, one workshop participant proposed dividing the blurred term ‘AI agent’ into three agent archetypes: reactive (lies dormant until a user triggers it to take on a task within a sandboxed environment), interactive (navigates multi-agent environments but continues to be triggered by a user to perform a task), and proactive (self-initiates tasks, proposes its own projects, and operates without constant human intervention in a multi-agent environment).[1]

The division could provide a flexible framework to align oversight with capabilities similar to Autonomy Safety Levels for autonomous vehicles.[2][3] Other useful conceptual work on the emerging characteristics of agentic AI systems has been recently conducted by the OECD.[4]

For the sake of clarity and brevity, this memo picks the following definitions as its foundation:

AI System

A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.[5]

AI Agent

AI agents are systems that can perceive and act upon their environment with a degree of autonomy, using tools as needed to achieve specific goals and adapt to changing inputs and contexts.[1]

Multi-Agent System

A setting in which agents interact with other agents – human or artificial. This can mean a single actor deploying a group of agents, but also multiple actors deploying different agents.

Oversight

We use oversight as an umbrella term spanning: (a) automated, embedded monitoring and containment that operates faster than humans can react, (b) human oversight, such as a designated human with the authority and the practical ability to understand, contest and change a system’s actions or outcomes,[6] which may be anticipatory (specified before deployment) or reactive (engaged at runtime); and (c) institutional or regulatory oversight by accountable public bodies.

Why Act Now

While many recent safety measures have focused on individual AI agents, multi-agent risks demand immediate and specific attention. There is a strong market incentive to build agentic capabilities. Multi-agent systems have been launched or prototyped to operate entire marketplaces,[7] scale procurement negotiations,[8] and take on enterprise workflows.[9] These systems are likely to grow and become more sophisticated soon. We urgently need to understand and prepare to mitigate the full spectrum of risks that may arise specifically from multi-agent behaviour because:

- **Safety does not carry over from single-agent to multi-agent settings.**[10][11] Even when individual agents pass safety audits, new risks emerge as soon as an agent is deployed in a network with other agents.[12] For example, AI systems in financial markets have demonstrated collusive behaviours,[13] and simulations have proven that different models can be combined to create malicious outputs without triggering their individual safety mechanisms.[14]

- **Interactions can scale into systemic failures.** The sheer speed and scale at which automated systems interact can trigger cascading effects, where small issues propagate across networks and become system-level vulnerabilities.[11] The 2010 'Flash Crash', where market prices collapsed by nearly a trillion dollars within minutes as automated trading systems reacted to each other, is a prime example.[15] Simulations have also shown how falsified data can lead to widespread outages in critical infrastructure such as power grids,[16] or resource depletion in simulated resource-sharing scenarios.[17]

- **Responsibility is fragmented across actors.** There is a divided value chain where different actors may be responsible for the underlying general-purpose AI model, the agent application built on it, and its real-world operation.[18] All of which are factors that can shape multi-agent risks during deployment.[18] This fragmentation requires every actor in the AI system lifecycle (Table 1) to implement specific multi-agent safety mechanisms, most of which are currently non-existent.[18]

- **Incident frequency is increasing.** The OECD AI Incidents Monitor (AIM)[19] reported a threefold increase in agentic AI incidents between November 2025 and February 2026. As AI agents become more capable and more widely deployed, their interactions may increasingly generate unintended and hard-to-predict effects, with potential implications for financial markets, digital infrastructure, and fundamental rights.

Luis Aranda
Senior Economist, OECD

‘We are witnessing a shift from AI systems that shape human decisions through generated content to increasingly agentic systems capable of more autonomous action. As a result, the risk frontier is moving from harmful content to autonomous actions – and ‘misactions’ – with potentially systemic effects.’

AI Actors Operating in the AI System Lifecycle [20]

More insight into possibilities and limitations of AI systems. Can provide insights into the potential risks, opportunities, and impacts of AI systems.[i]

More insight into real-world applications of AI systems in specific contexts. Can provide feedback on AI system performance and incidents.[i]

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

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.[ii]

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.[ii]

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.

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

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.

[i] Partnership on AI. 2026. ‘*Corporate AI Risk Assessment Framework*’ Partnership on AI. https://partnershiponai.org/moving-from-theory-to-action-in-ai-risk-management/.

[ii] European Parliament and Council of the European Union. 2024. ‘*Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*’ Official Journal of the European Union L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

What Industry and Third-Party Evaluators Can Do

Develop Multi-Agent Evaluation Methods

Today's AI systems are developed and primarily tested in isolation, despite the fact that they will soon interact with each other. In order to understand how likely and severe multi-agent risks are, providers, deployers, and third-party evaluators must develop evaluations and sandboxes that:

- **Measure tendencies to conflict and miscoordination.** Prior to interactive scenarios, individual agents must be tested to ensure they continue to choose socially beneficial actions in realistic situations where incentives encourage harmful behaviour. Benchmarks grounded in documented AI risk scenarios show frontier models often fail even at this stage.[21][22]

- **Measure robustness to adversarial influence.** Ensuring agents maintain intended behaviour when exposed to untrusted or manipulative inputs from other agents.[23]

- **Measure robustness to collusion.** This is particularly crucial since our most promising AI safety approaches, such as adversarial training and scalable oversight, rely on the assumption that the supervised and supervising systems have sufficiently divergent objectives or failure modes. Collusion, or correlated failures from shared training data and architectures, undermines this assumption, so multi-agent evaluations must probe for it directly.[11]

- **Measure the risks emerging specifically from multi-agent interactions,** such as manipulated knowledge cascading through a network of agents and causing infectious adversarial attacks,[24] or a group of agents and other systems developing emerging capabilities that go beyond that of each individual system.[25]

- **Factor in multiple providers.** Robust multi-provider sandbox testing should be developed to capture how different models, systems, and providers interact in practice, with repeated runs across diverse simulation scenarios.

Beyond technical performance, evaluations should also assess whether systems adhere to norms and regulations grounded in publicly accountable institutions. The legal alignment of agents can provide a structural blueprint for confronting challenges of reliability, trust, and cooperation in AI systems across jurisdictions.[26] Existing evaluations generally don't measure if agentic AI systems comply with relevant law across domains,[26] but promising first steps are being made with several benchmarks focusing on narrow regulatory contexts, such as the EU GDPR and EU AI Act.[27][28][29]

Additionally, providers must supply model documentation that explicitly addresses interaction-relevant characteristics, including behaviour under both competitive and cooperative conditions, communication patterns and protocols, the boundaries of agent action spaces, and any explicit caveats where multi-agent testing has not been conducted. This documentation should be extended to consider how other systems may interpret and act upon it in downstream contexts.[18]

Elham Tabassi
Director, AI and Emerging Technology Initiative and Senior Fellow, The Brookings Institution

'Prevailing evaluation methods largely measure task performance in controlled settings, while real deployments involve untrusted inputs, often absent identity verification, shifting interaction patterns, and network effects that single-agent benchmarks are not designed to capture.

Benchmarks like SWE-bench and WebArena are useful for assessing whether individual agents can complete defined objectives, such as resolving software issues or carrying out structured web tasks. These are single-agent task benchmarks. Emerging frameworks such as MultiAgentBench have begun to evaluate coordination and competition between multiple LLM-based agents across cooperative and adversarial scenarios, but remain at an early stage.'

Develop Technical Safety Solutions For Multi-Agent Environments

Even with preventive mechanisms such as evaluations and sandboxes in place, it may only become clear which agent dynamics require mitigation once a tipping point has been reached in deployment. Technical researchers at the workshop stressed the importance of tools to detect and respond to these behavioural tipping points before they cascade into systemic failures, something likely to require domain-specific approaches.

They primarily discussed:

- **Embedded containment and control.** Given the speed and scale of multi-agent interactions, oversight should be built into AI infrastructure itself – [30] capable of detecting anomalies and triggering interventions such as kill switches or circuit breakers that act faster than human oversight. In this way, supervisory AI architecture can oversee interactions between other AI agents to prevent network-level risks. Recent NIST work on automated tools designed to continuously read AI outputs against trusted information is a promising first step.[31] The triggers, thresholds, and safe states of these mechanisms must be specified, validated, and authorised by anticipatory human oversight,[32] and should be complemented by mechanisms that let agents halt in a safe state and return control to a human overseer or principal (reactive human oversight).
- **Incentives for responsible behaviour.** System design can also shape how agents behave. For example, shared scoring or reputation systems could reward cooperation and penalise harmful or disruptive actions across providers. This would help align incentives so that agents are more likely to act in ways that support overall system stability, rather than working at cross-purposes.[33][34]

Zhijing Jin
Research Scientist, Max Planck Institute for Intelligent Systems; Assistant Professor, University of Toronto; and co-founder of EuroSafeAI

‘To prevent large cascading risks caused by multi-agent AI, one idea is to limit the number of interactions an agent can issue per day (or other time periods), as well as the volume of resources they manage, which would be especially important to set ‘circuit breakers’ in financial markets and critical infrastructure such as energy grids and supply chains.

But hard constraints alone are unlikely to be enough. An important research direction is designing mechanisms that move agents away from behaviours maximising individual gain at the expense of collective stability, and toward settings where pursuing one’s own interest also supports system-wide welfare.

Mechanisms such as reputation systems, automated contracts, oversight AI agents that monitor for collusion, and coordination tools for resolving disputes, could provide a game-theoretic foundation to keep increasingly autonomous multi-agent ecosystems cooperative, robust, and aligned with societal interests.’

What Policy and Governance Can Do

Establish Visibility and Oversight Mechanisms

Deployers can track their own agents, but this visibility does not extend to other actors in the AI lifecycle. Oversight bodies must have sufficient, appropriately-scoped - and where necessary privileged and confidential - visibility into foundation-model provenance and system-level behaviour to attribute actions and intervene. Workshop attendees called for:

- Standardised and robust protocols for **agent identification** such as runtime-bound AI system IDs[35] and AI model registries[36] that can support incident reporting and attribution in high-risk and regulated deployments.

- Improved **data retention protocols** for AI agents. Activity logs should enable action attribution (audit trails,[37] and protocols)[38] to responsible actors, real-time monitoring and retrospective investigation.

- **Certification systems** could provide third-party verification of agent properties, such as what tools they have access to, their autonomy level, and their data handling practices, enabling actors to make informed decisions about interaction.[18]

However, visibility alone is insufficient. Governance frameworks should also assume incidents will most likely occur and mandate:

- Continuous **post-deployment monitoring** and anomaly detection, coupled with mechanisms for rapid intervention. This includes the ability to constrain, override, or revoke agent actions on short timescales (see section on embedded containment and control) and, when necessary, to shut down the system.[39][40]

- **Third-party reporting** mechanisms and interdisciplinary monitoring institutions. Because of the high-risk and diffused nature of multi-agent incidents, third-party reporting is indispensable to enable early detection of compounding effects.[41][42]

Julia Smakman
Senior Researcher in the Law and Policy Research Domain, Ada Lovelace Institute

‘Current legal frameworks (like the EU AI Act) focus on risk prevention for individual systems and models. This misses risks that emerge from the interaction between agents. We need processes that identify and log agent behaviour, particularly when they interact with other agents, to enable continuous risk monitoring. Without such measures, it is impossible for governance to intervene on time and prevent harm.’

Establish Incentive Mechanisms

The proprietary technical knowledge that private providers possess puts them in a better position to develop technical safety solutions such as embedded oversight or multi-provider sandboxing.[43] But applying such solutions requires enforcement or incentivisation from regulators. These private-public partnerships are currently non-existent between industry, policy and governance, but workshop participants suggested:

- Shifting incentives through **hard regulatory levers** such as liability frameworks and cross-provider evaluation mandates, which are powerful but slow.
- Introducing **market mechanisms** such as insurance,[43][44] procurement standards, and third-party auditors,[45] as well as investor pressure to accelerate adoption of safety practices, including schemes that require coverage for regulated systems and reward cross-provider cooperation at the system-level (e.g. mandatory activity logging between agents from different providers to enable external monitoring).

Apply Lessons From Adjacent Domains

Greater progress on multi-agent safety can be made by leveraging insights from fields that have already grappled with the challenge of coordinating multiple agents and stakeholders with potentially competing interests.

For example, AI agents are being released in a network of other agents with minimal regulatory frameworks following deployment. By contrast, one workshop participant highlighted how the automotive sector only permits autonomous vehicles onto networks of other vehicles and pedestrians after rigorous pre-deployment regulation and verification. Strong safety incentives have produced a system in which independent third parties verify safety claims and technical providers retain clear accountability, creating a slow but extremely risk-conscious innovation cycle.

While the automotive sector provides a cultural and procedural template, law already provides a normative backbone that can be leveraged for AI alignment and compliance across jurisdictions. Beyond generally enabling actors to cooperate without fear of counterparty defection or punishment,[46][47] private law rights could enable institutions, deployers, and AI systems to make credible commitments that promote strategic stability and safety.[48]

Conclusion

Current governance and risk management frameworks are not equipped to address system-level risks emerging from AI agents interacting at scale. Addressing these challenges will require a collaborative approach across all actors in the AI agent lifecycle.

Why We Should Care About Multi-Agent Risks

Agentic incidents are increasing.
Market incentives are stirring us towards a multi-agentic world.
If well governed, deployment can provide significant societal and economic benefits.

Multi-agent risks can scale rapidly into systemic failures.

Single-agent safety does not guarantee multi-agent safety.

Responsibility is fragmented across actors and multi-agent regulation is close to non-existent.

What To Do About It

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.

Identify

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

Monitor

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

Report

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

Incentivise

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

Notes

[1] Anwar, Usman. 2026. '*Three Archetypes of AI Agents — or Why Claude (Code) and Claudius Are Different Kinds of Agents*' Unaligned Thoughts (Substack). https://unalignedthoughts.substack.com/p/three-archetypes-of-ai-agents-or.

[2] SAE International. 2021. '*Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*' SAE Recommended Practice J3016_202104. https://doi.org/10.4271/J3016_202104.

[3] Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, et al. 2023. '*Harms from Increasingly Agentic Algorithmic Systems*' arXiv. https://arxiv.org/abs/2302.10329.

[4] OECD. 2026. '*The Agentic AI Landscape and Its Conceptual Foundations*' OECD Artificial Intelligence Papers, No. 56. Paris: OECD Publishing. https://doi.org/10.1787/396cf758-en.

[5] OECD. 2019. '*Recommendation of the Council on Artificial Intelligence*' OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[6] Sterz, Sarah, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. '*On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives*' https://doi.org/10.1145/3630106.3659051.

[7] Athropic. 2026. '*Project Deal*' 2026. Anthropic. 2026. https://www.anthropic.com/features/project-deal.

[8] Pactum AI. n.d. '*The Leading AI Procurement Negotiation Tool*' Pactum. https://pactum.com/.

[9] Higuchi, Takuto. 2025. '*Introducing Microsoft Agent Framework: The Open-Source Engine for Agentic AI Apps*' Microsoft Foundry Blog.

[10] Reid, Alistair, Simon O'Callaghan, Liam Carroll, and Tiberio Caetano. 2025. '*Risk Analysis Techniques for Governed LLM-Based Multi-Agent Systems*' arXiv. https://arxiv.org/abs/2508.05687.

[11] Hammond, Lewis, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, et al. 2025. '*Multi-Agent Risks from Advanced AI*' arXiv. https://arxiv.org/abs/2502.14143.

[12] Shapira, Natalie, Chris Wendler, Avery Yen, Gabriele Sarti, Koyena Pal, Olivia Floody, Adam Belfki, et al. 2026. '*Agents of Chaos*' arXiv. https://arxiv.org/abs/2602.20021.

[13] Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu. 2020. '*Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market*' Queen's Economics Department Working Paper No. 1438. Kingston, ON: Queen's University, Department of Economics.

[14] Jones, Erik, Anca Dragan, and Jacob Steinhardt. 2024. '*Adversaries Can Misuse Combinations of Safe Models*' arXiv. https://arxiv.org/abs/2406.14595.

[15] U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission. 2010. '*Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*' https://www.sec.gov/news/studies/2010/marketevents-report.pdf.

[16] Acharya, Samrat, Yury Dvorkin, and Ramesh Karri. 2021. '*Causative Cyberattacks on Online Learning-Based Automated Demand Response Systems*' IEEE Transactions on Smart Grid 12 (4): 3548–59. https://doi.org/10.1109/tsg.2021.3067896.

[17] Piatti, Giorgio, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. '*Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents*' arXiv. https://arxiv.org/abs/2404.16698.

[18] Nugteren, Maartje. 2026. '*Operationalizing the EU AI Act for Multi-Agent AI Systems.*' AI Policy Fellowship Publications. Montreal: Mila – Quebec AI Institute. https://mila.quebec/sites/default/files/media-library/pdf/614964/4maartje-nugteren-mila-ai-policy-fellowship-brief.pdf.

[19] OECD. 2024. '*OECD AI Policy Observatory Portal*' Oecd.ai. https://oecd.ai/en/incidents.

[20] This table builds on Partnership on AI. 2026. '*Corporate AI Risk Assessment Framework*' Partnership on AI. https://partnershiponai.org/moving-from-theory-to-action-in-ai-risk-management/.

[21] Cobben, Pepijn, Xuanqiang Angelo Huang, Thao Amelia Pham, Isabel Dahlgren, Terry Jingchen Zhang, and Zhijing Jin. 2026. '*GT-HarmBench: Benchmarking AI Safety Risks Through the Lens of Game Theory*' arXiv. https://arxiv.org/abs/2602.12316.

[22] Smith, Chandler, Cecilia Elena Tilli, Qi Guo, et al. Forthcoming. '*Inter-Agent Influence: Evaluating Persuasion, Deception and Coercion in Multi-Agent Systems.*'

[23] Franklin, Matija, Nenad Tomašev, Julian Jacobs, Joel Z. Leibo, and Simon Osindero. 2026. '*AI Agent Traps*' SSRN. https://doi.org/10.2139/ssrn.6372438.

[24] Ju, Tianjie, Yiting Wang, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, and Gongshen Liu. 2024. '*Flooding Spread of Manipulated Knowledge in LLM-Based Multi-Agent Communities'* arXiv. https://arxiv.org/abs/2407.07791.

[25] Jones, Erik, Anca Dragan, and Jacob Steinhardt. 2024. '*Adversaries Can Misuse Combinations of Safe Models*' arXiv. https://arxiv.org/abs/2406.14595.

[26] Kolt, Noam, Nicholas Caputo, Jack Boeglin, Cullen O'Keefe, Rishi Bommasani, Stephen Casper, Mariano-Florentino Cuéllar, et al. 2026. '*Legal Alignment for Safe and Ethical AI*' arXiv. https://arxiv.org/abs/2601.04175.

[27] Hu, Wenbin, Huihao Jing, Haochen Shi, Haoran Li, and Yangqiu Song. 2025. '*Safety Compliance: Rethinking LLM Safety Reasoning through the Lens of Compliance*' arXiv. https://arxiv.org/abs/2509.22250.

[28] Lichkovski, Ilija, Alexander Müller, Mariam Ibrahim, and Tiwai Mhundwa. 2025. '*EU-Agent-Bench: Measuring Illegal Behavior of LLM Agents under EU Law*' arXiv. https://arxiv.org/abs/2510.21524v1.

[29] Marino, Bill, Rosco Hunter, Christoph Schnabl, Zubair Jamali, Kalpakos, Marinos Emmanouil, Mudra Kashyap, Isaiah Hinton, et al. 2025. '*AIReg-Bench: Benchmarking Language Models That Assess AI Regulation Compliance*' arXiv. https://arxiv.org/abs/2510.01474.

[30] Chan, Alan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K Hadfield, and Markus Anderljung. 2025. '*Infrastructure for AI Agents.*' arXiv. 2025. https://arxiv.org/abs/2501.10114.

[31] NIST. 2026. '*Building Measurement Probes into Agentic AI Ecosystems*' NIST Information Technology Laboratory AI Webinar Series. https://www.nist.gov/news-events/events/2026/04/nist-information-technology-laboratory-ai-webinar-series-building.

[32] Gaube, Susanne, Markus Langer, Tim Miller, Kevin Baum, Raimund Dachselt, Anna Maria Feit, Ujwal Gadiraju, et al. 2026. '*Keeping an Eye on AI: A Framework for Effective Human Oversight of AI Systems*' arXiv. https://arxiv.org/abs/2605.16278.

[33] Piedrahita, David Guzman, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. 2025. '*Corrupted by Reasoning: Reasoning Language Models Become Free-Riders in Public Goods Games*' arXiv. https://arxiv.org/abs/2506.23276.

[34] Tewolde, Emanuel, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. 2026. '*CoopEval: Benchmarking Cooperation-Sustaining Mechanisms and LLM Agents in Social Dilemmas*' arXiv. https://arxiv.org/abs/2604.15267.

[35] Chan, Alan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. 2024. '*IDs for AI Systems*' arXiv. https://arxiv.org/abs/2406.12137.

[36] McKernon, Elliot, Gwyn Glasser, Deric Cheng, and Gillian Hadfield. 2024. '*AI Model Registries: A Foundational Tool for AI Governance*' arXiv. https://arxiv.org/abs/2410.09645.

[37] Kroll, Joshua A. 2021. '*Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems*' In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 758–771. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445937.

[38] Agentic AI Safety Community of Practice. 2025. '*Safer Agentic AI Foundations, Volume 2, Issue 3.*' Version 1.2. Edited by Nell Watson and Ali Hessami. https://www.saferagenticai.org.

[39] Madkour, Nada, Jessica Newman, Deepika Raman, Krystal Jackson, Evan R. Murphy, and Charlotte Yuan. 2026. '*Agentic AI Risk-Management Standards Profile*' Version 1.0. Berkeley: UC Berkeley, Center for Long-Term Cybersecurity, AI Security Initiative. February 2026.

[40] OECD. 2024. '*OECD AI Principles: Robustness, security and safety (Principle 1.4)*', https://oecd.ai/en/dashboards/ai-principles/P8.

[41] Brundage, Miles, Noemi Dreksler, Aidan Homewood, Sean McGregor, Patricia Paskov, Conrad Stosz, Girish Sastry, et al. 2026. '*Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies.*' arXiv. https://arxiv.org/abs/2601.11699.

[42] Frontier Model Forum. '*Information Sharing, Incident Reporting, and Incident Response for Frontier AI Risks - Frontier Model Forum.*' 2026. https://www.frontiermodelforum.org/issue-briefs/information-sharing-incident-reporting-and-incident-response-for-frontier-ai-risks/.

[43] Hadfield, Gillian K., and Andrew Freedman. 2026. '*What Markets Can and Can't Do for AI Governance*' The Architecture of AI Governance, by Fathom (Substack). https://fathomai.substack.com/p/what-markets-can-and-cant-do-for.

[44] Trout, Cristian, Rajiv Dattani, Rune Kvist, Dean Ball, Tom Zick, Ugur Ozer, Kevin Kalinich, et al. 2026. '*Underwriting AI Agents: The Blueprint for an AI Insurance Stack.*' arXiv. https://arxiv.org/abs/2607.11999

[45] Hadfield, Gillian K, and Jack Clark. 2023. '*Regulatory Markets: The Future of AI Governance.*' arXiv. https://arxiv.org/abs/2304.04914.

[46] North, Douglass C., John Joseph Wallis, and Barry R. Weingast. 2009. *Violence and Social Orders.* Cambridge: Cambridge University Press.

[47] Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity and Poverty.* New York: Crown Publishers.

[48] Goldstein, Simon, and Peter Salib. 2025. '*AI Rights for Economic Flourishing*' SSRN. https://ssrn.com/abstract=5353214.

Acknowledgements

The organisers would like to thank the speakers and panellists of the workshop Establishing Foundational Principles and Thresholds for Multi-Agent AI Governance, held in conjunction with IASEAI 2026 at UNESCO House, Paris, on 26 February 2026, for sharing their insights and expertise during the discussions that informed the development of this policy memo.

We are grateful to Elham Tabassi, Joel Z. Leibo, Julia Smakman, Kevin Baum, Luis Aranda, Merve Hickok, Patricia Paskov, Trish Shaw, and Zhijing Jin for their thoughtful comments and suggestions on earlier drafts of this memo.

Get in Touch
info@cooperativeai.org
www.cooperativeai.org

Connect
LinkedIn
X
Youtube