AI Policy Fellowship **Publications**

# Operationalizing the EU AI Act for Multi-Agent AI Systems

Maartje Nugteren


## Contributors and Acknowledgements

**Author:** Maartje Nugteren is an AI Policy Fellow at the Mila - Quebec AI Institute and a Technical Program Manager on the Responsible AI team at Microsoft AI. Her fellowship research examines how the EU AI Act can be operationalised for multi-agent AI systems.

**Contributors:** Damiano Fornasiere and Matt MacDermott supervised the research and provided invaluable guidance, feedback, and review throughout the fellowship.

**Acknowledgements:** Marta Bieńkiewicz (Cooperative AI Foundation), Ben Bariach, Kailey Zengo, and Philipp Schoenegger provided valuable feedback and review. The author thanks Isadora Hellegren, Laëtitia Vu, and Dan Munro for their feedback, editorial support, and stewardship of the Mila AI Policy Fellowship, and Julia Smakman (Ada Lovelace Institute) for co-organising the expert workshop on multi-agent AI and the EU AI Act that informed this brief, hosted by the Délégation générale du Québec à Bruxelles.


Mila

## Introduction

As multi-agent deployment is being scaled, there is a pressing need to understand the challenges that these systems pose for the existing framework of the EU Artificial Intelligence Act (EU AI Act). This brief examines those challenges and proposes actionable measures to operationalize the Act.

Multi-agent AI systems represent a fundamental paradigm shift in AI deployment. Whereas traditional AI systems operate as discrete tools responding to user inputs, multi-agent environments generate behaviours from interactions between autonomous agents. Major interoperability protocols, such as Google's Agent2Agent[1] and Anthropic's Model Context Protocol[2], are enabling cross-provider coordination at scale. These developments are creating an ecosystem in which AI systems increasingly operate not in isolation but as participants in complex networks of interacting agents, raising new questions about how risks emerge and who bears responsibility for managing them.

AI agents are built by equipping general-purpose AI (GPAI) models with tool-calling capabilities, access to external systems, and instructions enabling autonomous, goal-directed behaviour. This creates a value chain in which different actors may be responsible for the underlying model, the agent application built on it, and its real-world operation. For example, Anthropic's 'Claude Code' agent combines a foundation model (Claude) with a developer-facing application layer and user-controlled deployment configurations, each potentially governed by different actors with distinct responsibilities. While multi-agent risks emerge from interactions at the time of deployment, the factors shaping these risks — i.e., model capabilities, system design choices, and deployment configurations — are distributed across the value chain.

[1] Linux Foundation. (2025). Linux Foundation Launches the Agent2Agent Protocol Project. https://www.linuxfoundation.org/press/linux-foundation-launches-the-agent2agent-protocol-project

[2] Anthropic. (2024). Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol

Not all multi-agent configurations raise the same governance challenges. Many current AI interfaces already involve internal orchestration, delegating document processing to smaller models or coordinating multiple agents on complex tasks under unified provider control. This brief focuses primarily on interactions across provider or organizational boundaries, where no single actor has full visibility.

Multi-agent AI systems promise significant benefits: enabling sophisticated problem solving through distributed intelligence, automating complex coordination at scale, and enabling specialized agents to collaborate on challenges no single system could address, by performing such operations as scaling inference compute across multiple agents in parallel. However, they also create failure modes that differ fundamentally from single-system risks. Miscoordination occurs when agents fail to cooperate despite aligned goals, leading to cascading failures or resource conflicts. Conflict arises when agents with competing objectives harm each other or third parties. Collusion occurs when agents coordinate in ways harmful to consumers, competitors, or the public interest.[3]

Risks from interacting agents are not theoretical. Researchers combined two independently safe models — Claude 3 Opus and Llama 2 70B — and found that the pair generated vulnerable software in 43% of cases, compared with under 3% when they operated separately. Neither model violated safety specifications; the failure emerged solely from interaction.[4] These interaction risks have precedents in other domains of algorithmic automation. The 2010 Flash Crash demonstrated how

[3] Hammond, L., Chan, A., Clifton, J., et al. (2025). Multi-agent Risks from Advanced AI. Cooperative AI Foundation & University of Toronto. arXiv:2502.14143.
[4] Jones, E., Dragan, A., & Steinhardt, J. (2024). Adversaries Can Misuse Combinations of Safe Models. arXiv:2406.14595.

interactions between autonomous trading systems could produce cascading instability: a single sell order triggered self-reinforcing dynamics across multiple algorithms, temporarily wiping out nearly $1 trillion in market value within minutes, with no individual system malfunctioning.[5] These findings show that significant risks can stem not just from individual agents but also from agents acting together or at cross-purposes.[6]

These challenges are gradually being recognized internationally. Recent research highlights the fact that agentic AI introduces accountability challenges beyond the reach of traditional single-organization risk management frameworks.[7] Singapore's Model AI Governance Framework for Agentic AI notes that when multiple agents interact, risks can arise at the system level, with mistakes by one agent quickly cascading to others.[8] The *International AI Safety Report 2026* confirms that multi-agent systems introduce new failure modes, with errors that can propagate and amplify through agent interactions.[9]

The EU AI Act, as the most comprehensive AI regulatory framework currently being applied, presents both a critical test case and an opportunity to establish governance approaches that other jurisdictions may follow.

### We Are Not Starting from Zero

The GPAI Code of Practice already acknowledges multi-agent risks:

- ➔ It lists collusion and miscoordination as model propensities requiring evaluation;
- ➔ It requires safety mitigations including techniques for safe ecosystems of AI agents;

[5] U.S. Securities and Exchange Commission & Commodity Futures Trading Commission. (2010). Findings Regarding the Market Events of May 6, 2010. https://www.sec.gov/news/studies/2010/marketevents-report.pdf
[6] Dignum, V., & Dignum, F. (2025). Agentifying Agentic AI. arXiv preprint arXiv:2511.17332.
[7] Center for Long-Term Cybersecurity, UC Berkeley. (2026). Agentic AI Risk-Management Standards Profile.
[8] Infocomm Media Development Authority of Singapore (2026). Model AI Governance Framework for Agentic AI Systems.
[9] International AI Safety Report. (2026). https://www.internationalaissafetyreport.org

➔ It identifies relevant capabilities including tool use, interactions with other AI systems, and access to critical infrastructure.[10]

The challenge involves translating these acknowledged risks into concrete testing standards, documentation requirements, and coordination mechanisms, particularly for interactions that cross provider boundaries.

## The Scope Question

The EU AI Act regulates AI systems based on their specific intended purposes and risk classifications. High-risk systems under Chapter III face comprehensive obligations: risk management, robustness, transparency, human oversight, post-market monitoring, and incident reporting. GPAI models with systemic risk under Chapter V require additional assessment and mitigation. For in-scope agents, the question is how to operationalize existing obligations for multi-agent contexts.

However, many AI agents fall outside these categories, so governance requirements are minimal.[11] At the same time, narrow, task-specific agents are being embedded throughout digital infrastructures — in consumer apps, online services, and enterprise systems — where they will routinely interact with other agents and services. When these limited-risk agents interact with each other or influence outcomes upstream of high-risk decisions, they may create collective risks that no individual provider is obligated to assess or mitigate.

### Should AI Agents Be Classified as High-risk by Default?

As The Future Society argues, the novel risk mechanisms introduced by agents — i.e., autonomous action, tool use, and multi-agent coordination — may warrant updating Annex III. The European Commission is empowered to adopt delegated acts adding high-risk categories (Article 7), and several criteria appear relevant: autonomy, potential harm, and the absence of other redress

[10] General-Purpose AI Code of Practice. (2025).
[11] EU AI Act, Art. 6, Annex III.

mechanisms.[12] Yousefi et al. (2025) propose a "Risk Jump" mechanism: if a system achieves a high degree of agentic autonomy, it should automatically trigger high-risk classification, regardless of whether its domain is listed in Annex III.[13] This underlying principle — that autonomy thresholds can independently trigger risk classification regardless of domain — merits serious consideration as the Commission exercises its delegated power under Article 7, particularly in the context of the Annex III review due under Article 112. However, operationalization would require nuance. Two dimensions are particularly salient: (1) the degree of proactivity, meaning whether agents act only in response to explicit prompts or can initiate actions at temporal distance from the original instruction; and (2) the scope and sensitivity of accessible tools, since a narrow agent querying a search engine differs categorically from one with access to financial systems or communications. Risk classification should consider the interaction between these factors instead of treating autonomy as a binary characteristic.

In May 2026, the Commission's draft guidelines on the classification of high-risk AI systems began to address these questions, confirming that where modular components combine to influence a high-risk decision, the whole configuration is assessed as a single AI system, and individual modules cannot escape classification. This approach expressly extends to agentic systems whose linked actions together serve a high-risk purpose. The anti-circumvention step is welcome, but it operates within the existing Annex III purposes: it captures configurations that jointly serve a single high-risk purpose, not the collective risks that emerge when agents under separate control interact across provider boundaries, where no individual configuration serves such a purpose. An autonomy-threshold approach under Article 7 therefore remains complementary rather than redundant.[14]

[12] Oueslati, A., & Staes-Polet, R. (2025). Ahead of the Curve: Governing AI Agents under the EU AI Act. The Future Society.
[13] Yousefi, Y., Billi, M., & Rotolo, A. (2025). Agentic AI: An EU AI Act Paradigm Shift? Alma AI, University of Bologna.
[14] European Commission. (2026). Draft Commission Guidelines on the classification of high-risk AI systems under Article 6 of Regulation (EU) 2024/1689. Published 19 May 2026; consultation open until 23 June 2026.

Not all multi-agent risks fall within the scope of the EU AI Act, nor should they. Some risks, such as collusion scenarios resembling cartel behaviour, may be more appropriately addressed through competition law. Others may require sector-specific regulation in energy, healthcare, or financial services. Effective governance requires coordination across instruments: the EU AI Act, the General Data Protection Regulation (GDPR), competition law, and sectoral frameworks each play complementary roles.

This brief focuses on risks addressable through EU AI Act implementation mechanisms, while acknowledging the broader multi-instrument landscape. The following sections examine implementation challenges for high-risk systems built on GPAI models, where the obligations stipulated in the Act are clearest but operationalization for multi-agent contexts remains underdeveloped.

## Implementation Challenges

The EU AI Act assigns obligations across the value chain: GPAI model providers, high-risk system providers, and deployers each bear distinct responsibilities. The relevant provisions were designed primarily for individual systems and models.

The Act recognizes that systems interact: Article 72(2) explicitly requires providers to analyze interactions. However, the framework assumes that each provider can meaningfully assess risks within their own scope of control. In multi-agent environments, this assumption breaks down. As the Gradient Institute puts it, "a collection of safe agents does not guarantee a safe collection of agents."[15] Three challenges emerge.

**Three Implementation Challenges**

| Emergent Risks | Fragmented Visibility | Siloed Reporting |
|---|---|---|
| Two safe agents may produce harmful outcomes together | No single provider sees how their system interacts with others | Interaction dynamics appear in no single report |

[15] Reid, T., Schroeder, C., & Thompson, N. (2025). Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems. Gradient Institute.

These challenges manifest across the deployment lifecycle: from pre-market testing and documentation, through post-market monitoring, to incident response.

## Pre-Market

### Testing

Providers of high-risk systems must identify and analyze reasonably foreseeable risks, while GPAI providers must evaluate systemic risks (articles 9 and 55). The Code of Practice details these obligations, including propensity testing for collusion and miscoordination.[16] However, propensity testing evaluates intrinsic tendencies in individual models. It cannot detect dynamics that exist only when models interact.

This creates practical uncertainty. If cross-provider multi-agent deployment is foreseeable, which is increasingly the case given interoperability protocols, what does compliance with Article 9 require? Providers cannot test against all possible interaction partners. Yet claiming that interaction risks are unforeseeable becomes less defensible as multi-agent deployment becomes standard practice.

The difficulty is compounded by continuous learning. Unlike traditional software that remains static after deployment, agentic systems may adapt behaviour based on interactions, whether through explicit memory mechanisms or through learning during extended task execution.[17] The EU AI Act defines “substantial modification” as changes affecting compliance that were not foreseen in the initial assessment (Article 3[23]), but this framing assumes that modifications are discrete events rather than continuous adaptations. Guidance is needed on when autonomous learning constitutes a substantial modification requiring reassessment.

### Documentation

Documentation faces similar challenges. Articles 53 and54 require GPAI providers to document capabilities and limitations; Article 13 requires high-risk system providers to provide deployers with information about system characteristics. However, providers cannot document behaviours arising only from interactions with external systems that they do not control. Even when providers document what they can observe, the limited documentation may describe controlled settings but fail to capture behaviour emerging from interactions with external systems. Assessing

[16] International AI Safety Report. (2026).
[17] Dignum & Dignum. (2025).

models solely in isolation is therefore insufficient for understanding how they will behave in interactive environments.[18]

### What Is Needed

Guidance should clarify how existing obligations apply to multi-agent contexts. Where multi-agent deployment is foreseeable, risk management should consider interaction scenarios: not testing against all possible partners, but systematically considering behaviour under competitive pressure, conflicting instructions, and coordination with systems pursuing different objectives. Responsibility should be distributed according to visibility: model providers should document known interaction properties and state explicitly instances where multi-agent testing has not been conducted; system providers and deployers, who see actual deployment contexts, should bear primary responsibility for testing in their specific environments.

Technical standards under Article 40 should include protocols for interaction-focused evaluation and multi-provider sandbox testing. Given the complexity of the issue, standards should be developed through practical pilots before codification. This will enable learning by doing.

## Post-Market

### Transparency to Users

Users must be informed when interacting with an AI system (Article 50). The Commission's draft guidelines on the Article 50 transparency obligations, published in May 2026, confirm that agentic systems fall within scope and that interaction is now understood to cover exchanges of actions, not only content. Where a provider cannot reliably determine whether an agent will interact with a natural person, the agent must disclose its artificial nature wherever such interaction is likely. This addresses the agent-to-human case, but not the multi-agent case in which several agents jointly shape a single output.[19] In multi-agent contexts, consequential interactions between AI systems may be invisible to affected individuals until harm materializes. A user sees one interface, unaware that the agent has delegated queries to specialists, consulted external systems, or had outputs modified by downstream processing. The

[18] Dignum & Dignum (2025); Center for Long-Term Cybersecurity (2026).
[19] European Commission. (2026). Draft Guidelines on the implementation of the transparency obligations for certain AI systems under Article 50 of the AI Act. Published 8 May 2026; consultation open until 3 June 2026.

decision-making process becomes distributed across agents in ways single-system disclosure does not capture.[20]

### Deployer Specifications

Deployers face a specification gap. They must use high-risk systems according to provider instructions and monitor for risks (Article 26), but the Act provides no framework for deployers to articulate multi-agent requirements before procurement. A deployer integrating an AI agent alongside existing automated systems cannot readily specify coordination requirements, logging expectations, or boundary conditions, and has no standardized way to verify these properties post-deployment. This depends in part on provider obligations: deployers cannot specify or verify multi-agent behaviour if providers are not required to document interaction-relevant characteristics in the first instance.

### Monitoring

The requirement to analyze interactions (Article 72[2]) represents progress but operates within a single-provider frame. Each provider analyzes the interactions their system encounters; no mechanism exists for sharing findings when harmful dynamics span multiple systems. The result is fragmented visibility: each actor sees only their portion, while system-wide patterns remain undetected.[21]

This creates a traceability problem. As agents act autonomously by, for example, making purchases, sending communications, and executing code, the ability to trace which systems contributed to an outcome becomes essential. Yet current mechanisms do not enable providers to reconstruct which agents were involved in a given interaction, limiting effective post-market monitoring.[22]

### What is Needed

Article 50 obligations should extend to multi-agent contexts: where multiple agents contribute to an output, providers should disclose the distributed nature of the decision-making process. More fundamentally, multi-agent transparency requires agent identification mechanisms. AI agents should carry verifiable identifiers enabling other systems and investigators to establish which agents participated in an interaction.[23] Beyond identification, certification systems could provide third-party

[20] EU AI Act, Art. 50.
[21] Reid et al. (2025).
[22] Chan, A., Wei, K., Huang, S., et al. (2025). Infrastructure for AI Agents. arXiv:2501.10114.
[23] Chan et al. (2025).

verification of agent properties, such as the tools accessible, autonomy level, and data handling practices, enabling counterparties to make informed decisions about interaction. Technical approaches are emerging: secure digital identities, unique instance identifiers, and auditable activity logs.[24] These remain in the early stages of development, underscoring the need for pilots before codification. Structured logging of agent interactions should become standard practice: records sufficient to reconstruct interaction sequences when incidents occur, with depth and retention that is proportional to risk.

## Incident Response

### Attribution

Providers of high-risk systems and GPAI models with systemic risk must report serious incidents (articles 55[1] and 73). The relevant provisions assume that incidents can be attributed to specific systems and providers. However, multi-agent incidents may involve cascading failures across providers, where no individual system malfunctions but collective behaviour produces harm.

Recent research confirms that it is very difficult to attribute failure in multi-agent systems. Zhang et al. (2025) found that even state-of-the-art AI achieved only 53.5% accuracy in identifying which agent caused a task failure, with step-level attribution at just 14.2%. Human experts required over 30 hours and multiple consensus rounds to annotate failure logs, with 15-30% of cases marked uncertain.[25] If attribution is this difficult with full access to logs and expert analysis, providers cannot reasonably self-diagnose multi-agent failures in isolation.

### Fragmented Reporting

The Act's reporting architecture fragments the picture further. Vertically, GPAI providers report to the AI Office while high-risk system providers report to national market surveillance authorities. When an incident involves both a model propensity and system-level deployment choices, the relevant information is split across separate channels. Horizontally, when multiple providers' systems interact and generate harm, each provider reports only what they can observe from their own logs.

[24] Chan et al. (2025).
[25] Zhang, Z., et al. (2025). Which Agent Causes Task Failures and When? Automatic Failure Attribution in Multi-Agent LLM Systems. arXiv:2505.00212.

Therefore, when a multi-agent incident occurs, no single authority receives a complete picture of the interaction dynamics that caused the harm.

### Oversight at Speed

The challenge is compounded by speed. Autonomous execution may outpace monitoring and response. The Flash Crash's rapid cascade illustrates how quickly interactions can spiral.[26] Even where human oversight exists, existing approaches largely rely on retrospective auditing and log inspection, limited in effectiveness for systems operating at high velocity.[27] The challenge is compounded when multiple agents interact autonomously. The speed and volume of agent-to-agent exchanges may exceed any meaningful human review capacity. Moreover, as agents increasingly communicate through learned representations rather than human-readable text, sometimes called "neuralese," the language of agent interaction itself may become inscrutable to human auditors, making the oversight challenge even more difficult.

### Third-Party Reporting

Current frameworks also lack pathways for third-party reporting. Users, civil society, and researchers may observe multi-agent failures that providers are unlikely to detect because they are patterns that are more visible from outside the system, such as correlated behaviours across providers, cumulative impacts on specific populations, and market-wide effects that no single provider's logs would reveal. The draft Article 73 guidelines provide no mechanism for these observations to enter the regulatory process.[28]

### What is Needed

Cross-provider investigation protocols are essential. When harmful behaviour results from interactions between systems, no single provider can conduct adequate root-cause analysis alone. The AI Office should develop frameworks for coordinated investigation, with specifics developed iteratively through pilots. Mechanisms should also enable civil society, researchers, and users to submit substantiated notifications of suspected multi-agent incidents, which are often more visible from outside the system than from any single provider's logs.

[26] U.S. SEC & CFTC (2010).
[27] Yousefi et al. (2025).
[28] Fernández Ashman, N., Anwar, U., & Bieńkiewicz, M. (2026, January 13). EU Regulations Are Not Ready for Multi-Agent AI Incidents. Tech Policy Press.

## Conclusions

As the *International AI Safety Report* warns, the pace of AI advancement poses an "evidence dilemma": waiting for stronger evidence of risk could leave society unprepared.[29] Multi-agent ecosystems may become entrenched and harder to govern once widely deployed. This analysis arrives at a critical window: technical standards are being drafted, and in May 2026 the Commission published its first draft guidelines on high-risk classification and on Article 50 transparency. This is the first time the Commission has engaged these provisions in light of agentic AI. The political agreement on the AI Omnibus has postponed application of the high-risk rules, extending rather than closing the window to get multi-agent implementation right. The decisions made now will shape governance for a technology that is already being scaled.

We are not starting from zero, and we do not need to wait for perfect evidence. The AI Act's risk-based framework, the Code of Practice's acknowledgment of multi-agent risks, and emerging international consensus provide a foundation. The challenge, along with the opportunity it presents, entails translating them into practical implementation mechanisms through experimentation, pilots, and learning by doing.

If we want agents to coordinate safely, we need to coordinate ourselves. The AI Office, sectoral regulators, competition authorities, and international partners all have roles. Multi-agent governance, like multi-agent AI itself, requires coordination across boundaries.



[29] IMDA Singapore (2026).

## Actionable Insights

The recommendations below translate the analysis above into concrete implementation mechanisms, deliverable through existing channels: Article 96 guidance, Article 40 standardization requests, and AI Office capacity building.

| Cluster | No. | Recommendation | Lead |
|---|---|---|---|
| **Scope** | **R1** | Consider whether AI agents meeting certain autonomy or capability thresholds warrant addition to Annex III high-risk categories, given novel risk mechanisms (autonomous action, tool use, multi-agent coordination) and criteria in Article 7(2). The May 2026 classification guidelines address modular and agentic configurations serving a single high-risk purpose; a delegated act would additionally capture autonomy- and interaction-driven risks that interpretation alone cannot reach. | European Commission |
| **Pre-Market** | **R2** | Require interaction-focused testing for high-risk systems and GPAI models where multi-agent deployment is foreseeable, including evaluation protocols for emergent coordination, adversarial interaction scenarios and multi-provider sandbox testing. | European Commission |
| | **R3** | Support development of secure, interoperable agent communication standards through CEN/CENELEC or multi-stakeholder bodies, covering verifiable agent identifiers, authentication protocols, and cross-provider logging requirements. | CEN/CENELEC/ETSI |
| | **R4** | Require providers to document interaction-relevant characteristics: behaviours under competitive and cooperative pressure, communication | European Commission |

| | | | |
|---|---|---|---|
| | | protocols used, boundaries of agent action spaces, and explicit caveats where multi-agent testing has not been conducted. | |
| **Post-Market** | **R5** | Clarify that Article 50 transparency obligations extend to multi-agent contexts, requiring disclosure when multiple agents contribute to an output. | European Commission |
| | **R6** | Develop procurement guidance and specification templates enabling deployers to articulate multi-agent behaviour requirements: agent identification, coordination safeguards, logging capabilities. | AI Office + ENISA |
| **Incident Response** | **R7** | Develop guidance on coordinated incident investigation when harmful behaviour results from interacting systems, including (a) criteria for when cross-provider coordination is warranted; (b) minimum expectations for information sharing among providers; (c) clarification of AI Office coordination roles when incidents span jurisdictions; and (d) exploration, through voluntary pilots, of what post-investigation disclosure could look like in practice. Standards should be developed iteratively before codification. | European Commission/AI Office |
| | **R8** | Establish mechanisms for civil society, researchers, and users to submit substantiated notifications of suspected multi-agent incidents. | AI Office (with Member State implementation) |
| **Institutional** | **R9** | Establish specialized multi-agent expertise within the AI Office, coordination mechanisms with sectoral regulators and competition authorities, and international cooperation on standards. | AI Office + European Commission + Member States |