June 2025

THE FUTURE SOCIETY

Ahead of the Curve

Governing AI Agents Under the EU AI Act

Amin Oueslati and
Robin Staes-Polet

Ahead of the Curve: Governing AI Agents under the EU AI Act

Contact: info@thefuturesociety.org

Cite as:
Amin Oueslati and Robin Staes-Polet, “Ahead of the Curve: Governing AI Agents under the EU AI Act” (The Future Society, June 2025).

Table of Contents

Executive Summary

This report provides the first comprehensive analysis of how the EU AI Act regulates AI agents, increasingly autonomous AI systems that can directly impact real-world environments. Our three primary findings are:

1. **The AI Act imposes requirements on the general-purpose AI (GPAI) models underlying AI agents (Ch. V) and the agent systems themselves (Ch. III).** We assume most agents rely on GPAI models with systemic risk (GPAISR). Accordingly, the applicability of various AI Act provisions depends on (a) whether agents proliferate systemic risks under Ch. V (Art. 55), and (b) whether they can be classified as high-risk systems under Ch. III. We find that (a) generally holds, requiring providers of GPAISRs to assess and mitigate systemic risks from AI agents. However, it is less clear whether AI agents will in all cases qualify as (b) high-risk AI systems, as this depends on the agent's specific use case. When built on GPAI models, AI agents should be considered high-risk GPAI systems, unless the GPAI model provider deliberately excluded high-risk uses from the intended purposes for which the model may be used.
2. **Managing agent risks effectively requires governance along the entire value chain.** The governance of AI agents illustrates the "many hands problem", where accountability is obscured due to the unclear allocation of responsibility across a multi-stakeholder value chain. We show how requirements must be distributed along the value chain, accounting for the various asymmetries between actors, such as the superior resources and expertise of model providers and the context-specific information available to downstream system providers and deployers. In general, model providers must build the fundamental infrastructure, system providers must adapt these tools to their specific contexts, and deployers must adhere to and apply these rules during operation.
3. **The AI Act governs AI agents through four primary pillars: risk assessment, transparency tools, technical deployment controls, and human oversight**. We derive these complementary pillars by conducting an integrative review of the AI governance literature and mapping the results onto the EU AI Act. Underlying these pillars, we identify 10 sub-measures for which we note specific requirements along the value chain, presenting an interdependent view of the obligations on GPAISR providers, system providers, and system deployers.

The report begins by examining major trends shaping the AI agent landscape. Numerous stakeholders, including the executives of major technology companies, predict that agents will have transformative societal impact. This aligns with recent acceleration in AI agent releases and improvements in their capabilities. Nonetheless, current agents are largely constrained to virtual environments, focusing on areas like software

development and computer use, whereas performance in other domains remains unreliable and mostly below human-level proficiency.

Absent a canonical definition, AI agents are best characterised through their functional properties and technical composition. The term “agent” is applied loosely within the tech industry, often neglecting the concept’s long and complicated history which dates back to the 1940s. With regards to current AI agents, two aspects stand out: functionally, current agents are defined by their capacity to autonomously pursue complex goals and take actions in both virtual and real-world environments; technically, they consist of a GPAI model integrated with auxiliary "scaffolding" such as chain-of-thought reasoning frameworks and tool access.

We identify autonomous long-term planning and direct real- and virtual-world interactions as key sources of risk from AI agents. These capabilities amplify risks under both Ch. III (high-risk applications) and Ch. V (systemic risks from GPAISRs) of the EU AI Act, creating new pathways to harm within existing domains–from multi-agent collusion in financial systems to large-scale manipulation.

Whereas it seems clear that Ch. V governs agent risks at the model-level, their classification as high-risk systems under Ch. III is less clear. Providers of GPAISRs must assess and mitigate systemic risks potentially arising from their integration in agent systems, whether the agent is developed by a downstream provider or by themselves. Under Ch. III, agents are considered high-risk if intended for use as a safety component (e.g., in medical devices, industrial machinery, or cars) or for a high-risk use case as specified in Annex III. Classification proves challenging, however, given that AI agents are potentially general-purpose systems, and could be arguably high-risk systems by default in the absence of deliberate exclusion of high-risk use by the provider. Moreover, the high-risk uses defined in Annex III predate current awareness of AI agent risks, raising questions about the adequacy of the current classification criteria.

Governing AI agents requires an effective distribution of obligations along the entire value chain. The “many hands problem,” which characterises value chains involving many actors, poses distinct challenges to accountability and risk management. With regards to AI agents, governance requirements must be divided among model providers, system providers, and system deployers, addressing asymmetries in expertise, resources, and information access. For instance, to monitor an AI agent, the model provider must set up a configurable monitoring infrastructure, the system provider sets thresholds for monitoring alerts, accounting for the intended use, and the system deployer oversees the agent monitoring during actual deployment.

We conduct an integrative literature review of the agent governance literature and **derive four complementary pillars**: **risk assessment, transparency tools, technical**

deployment controls, and human oversight. Mapping these pillars onto the requirements from the AI Act, we identify specific obligations associated with each measure for GPAISR providers, system providers, and system deployers. The resulting taxonomy of agent governance under the EU AI Act is presented in the table below. A more detailed discussion of these measures, along with specific references to the relevant provisions of the AI Act, is provided in Section 4.

Allocation of Governance Obligations Across Actors in the AI Agent Value Chain

Pillar	Measure	(GPAISR) Model Provider	(High-risk) System Provider	(High-risk) System Deployer
4.1 Agent Risk Assessments	4.1.1 Risk Identification	Map broad, high-level pathways to harm from agentic capabilities	Develop detailed risk scenarios specific to deployment context	Use the AI Office template to conduct fundamental rights impact assessments
	4.1.2 Risk Evaluation	Conduct general capability testing via standardised benchmarks	Perform continuous & iterative use-case specific testing reflecting operational conditions	Conduct fundamental rights impact assessments & use risk evaluation tools to guide go/no-go decisions
4.2 Transparency Tools	4.2.1 Agent Identifiers	Build core agent identification infrastructure & provide agent card templates	Implement & maintain agent IDs for specific deployment context	Ensure agent identification remains intact, accurate & recorded
	4.2.2 Real-Time Monitoring	Develop configurable monitoring capabilities & set default alert thresholds	Set context-appropriate thresholds & response protocols, provide feedback on alert accuracy	Monitor the agent, report risks & suspend use (if needed)
	4.2.3 Activity Logs	Develop & maintain detailed logging infrastructure for agent activity	Implement logging with deployment-specific detail & retention policies	Retain automatically generated logs
	4.2.4 Acceptable Use Policies (AUPs)	Define broad boundaries for agent use & key constraints	Review & adhere to model provider AUPs, develop supplementary policies for specific use cases	Follow both model & system providers' AUPs
4.3 Technical Deployment Controls	4.3.1 Real-Time Action Refusal	Build multi-level filtering frameworks for agent outputs	Add domain-specific filters & monitor effectiveness	Review filter performance, report issues & refine mitigation
	4.3.2 Emergency Shutdowns	Create automated & manual shutdown infrastructure linked to monitoring	Define graduated response protocols & investigation procedures	Test shutdown protocols, train staff & log incidents
4.4 Human Oversight	4.4.1 Checkpoint System	Develop configurable checkpoint mechanisms, linked to logging & emergency shutdown	Place checkpoints strategically & establish review protocols	Ensure their staff have the requisite AI literacy to make informed & less biased decisions
	4.4.2 Permission Management	Build permission control infrastructure, develop documentation of permissions	Configure granular permissions for specific operational needs & risks	Review permission configurations based on operational experience

Glossary

Note: The definitions provided below are drawn from the AI Act and other sources (such as the OECD). They are not strictly legal definitions, but are intended to clarify key concepts used in this report for better understanding and context.

Group	Term	Definition
Software	**AI agent**	Functionally, a system that autonomously pursues complex and long-term goals and takes actions in virtual and real-world environments. Technically, a compound system consisting of a general-purpose AI model (GPAI) and scaffolding.
Software	**AI system**	A machine-based system that is designed to operate with varying levels of autonomy, infer from input, and generate outputs that can influence environments.
Software	**General-purpose AI (GPAI) model**	An AI model trained with vast data using self-supervision at scale, exhibiting significant generality and capable of competently performing a wide range of tasks, and that can be integrated into a variety of downstream systems or applications.
Software	**GPAI model with systemic risk (GPAISR)**	A general-purpose AI model with systemic risk is a GPAI model with scale, capabilities, or market impact that presents significant potential risks to society, the economy, or fundamental rights, requiring stricter regulatory oversight. A GPAI model is presumed to carry high-impact capabilities when its training involves more than 10^{25} floating point operations (FLOPs).
Software	**GPAI system**	An AI system based on a general-purpose AI model with the capability to serve multiple purposes, either directly or when integrated into other systems.
Software	**Scaffolding**	External tools or software components integrated with a GPAI model to improve task performance, such as reasoning frameworks, memory, or tool access.
Features	**Agency**	The degree to which an AI system acts autonomously to pursue complex and long-term goals and takes actions in virtual and real-world environments.

Features	**Autonomy**	An AI system's ability to operate independently, making decisions and adapting to changing conditions with minimal human intervention.
Features	**Capability**	The proficiency level at which an AI system can perform tasks of varying difficulty.
Features	**Generality**	The breadth of tasks an AI system can perform competently, indicating its versatility across different domains.
Actors	**GPAI model provider**	An entity that develops and offers GPAI models. In the context of agents, these will commonly be GPAI models with systemic risk.
Actors	**System provider**	An entity or individual that develops, markets, or deploys an AI system under its own name or trademark, regardless of whether it is for profit or free.
Actors	**System deployer**	An entity using an AI system under its authority, except for personal, non-professional use.

1. State of AI Agents

“2025 is the year of AI agents”, claimed OpenAI Chief Product Officer Kevin Weil during this year’s World Economic Forum in Davos (Axios, 2025). With similar pomp, Salesforce CEO Marc Benioff announced in September 2024 that the goal of his company was to “empower one billion agents [...] by the end of 2025” (Salesforce, 2024). Predictions highlighting agents’ transformative effects in the near future were made by most major tech CEOs, including Sam Altman (2025) and Mark Zuckerberg (ZDNET, 2025).

Agents are AI applications that can execute complex tasks on their own, both in virtual and real-world settings. As of now, agents remain focused on specific tasks like managing inboxes and booking flights. But if they live up to their promise, they may soon become highly-capable digital coworkers or personal assistants. As such, the autonomy of agents, and their capacity to take direct action mark a fundamental shift, moving AI “through the chat window and into the real world” (CSET, 2025). While agents’ precise impact is yet to be seen, one thing is clear: major technology companies are making a concerted effort to turn their predictions into reality, investing large sums and releasing an ever-growing array of AI agent products and features.

1.1 AI Agent Trends

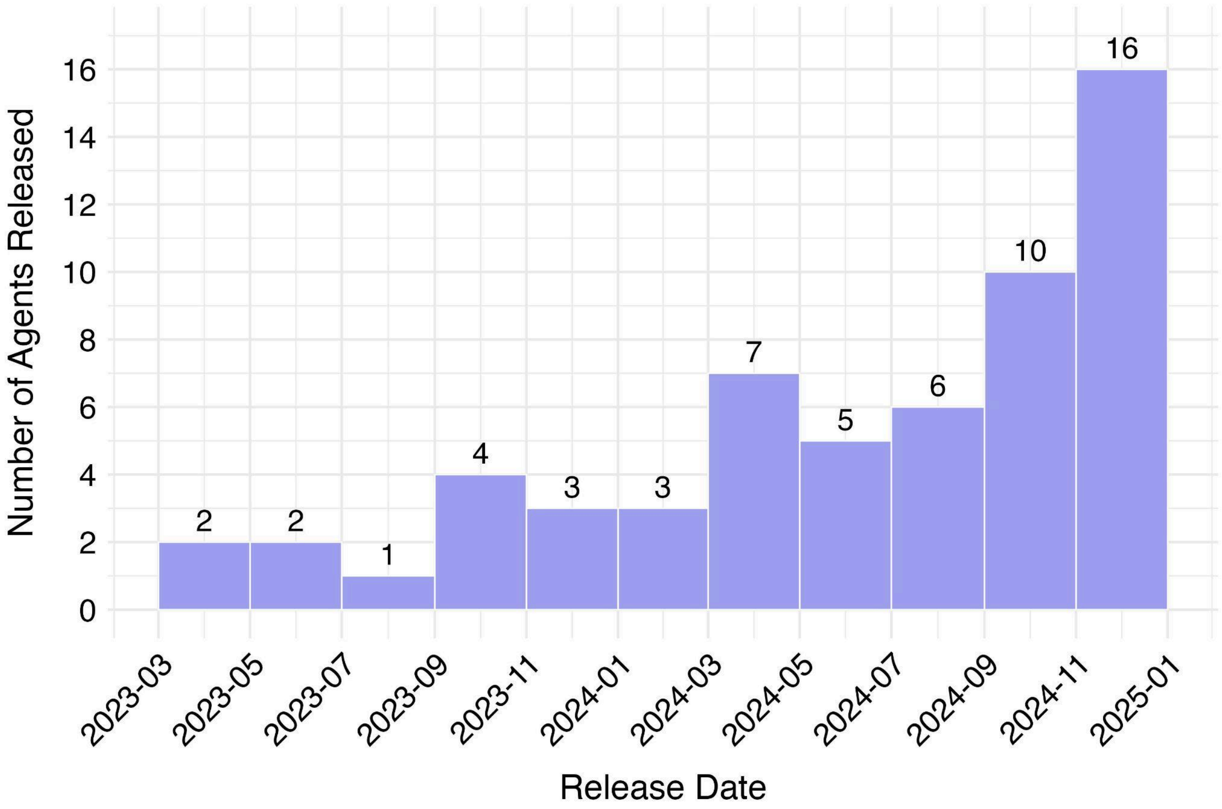

Figure 1. Number of agents released over bi-monthly time intervals between March 2023 and January 2025. Based on known release dates of AI agents included in the AI Agent Index developed by Casper et al. (2025).

AI agent deployment is accelerating. Casper et al.'s (2025) AI Agent Index gathers public information on released AI agents. According to the index, approximately half of all included agents were released during the second half of 2024. It is worth noting that other sources report much larger numbers. This includes the AI Agents Directory, a marketplace for AI agents, which contains more than 1,300 agents (AI Agents Marketplace & Directory, 2025). One explanation for the discrepancy lies in the AI Agent Index's more stringent inclusion criteria: whereas any organisation can submit their agent to the AI Agents Directory, Casper et al. (2025) require agents to demonstrate high degrees of agency and consequential impact, among other criteria. Apart from showcasing scholarly rigour, it exposes the current "hype", compelling companies to brand their offerings as AI agents.

In addition to smaller downstream providers, all major providers of GPAI models with systemic risk (GPAISR) are developing agents. Table 1 presents select examples of recent agent releases from major model providers.

GPAISR Provider	**Release Date**	**AI Agent Details**
Anthropic	October 2024	"**Computer Use**" is introduced to Claude, allowing it to control a user's computer interface (Anthropic, 2024b).
Google DeepMind	December 2024	"**Project Mariner**" is designed to solve open-ended web tasks (Google, 2024).
OpenAI	January 2025	"**Operator**" can perform web tasks for the user, such as booking travel, by using its own browser to access websites (OpenAI, 2025a).
OpenAI	February 2025	"**Deep Research**" enables multi-step research on the internet (OpenAI, 2025b).

Table 1. Non-exhaustive overview of agents released by major GPAISR providers.

Current agents remain primarily constrained to the virtual environment. Researchers and developers are building agents for a variety of domains, including research (Bran et al., 2023), software engineering (Jimenez et al., 2023), personal assistance (A. Chan et al., 2024), and cyber-offence (Fang et al., 2024), among others. Among the 67 systems included in the AI Agent Index, 74.6% either relate to computer use or software (Casper et al., 2025). The current concentration in virtual environments likely relates to the greater technical feasibility of virtual agents, among other factors (Toner et al., 2024).

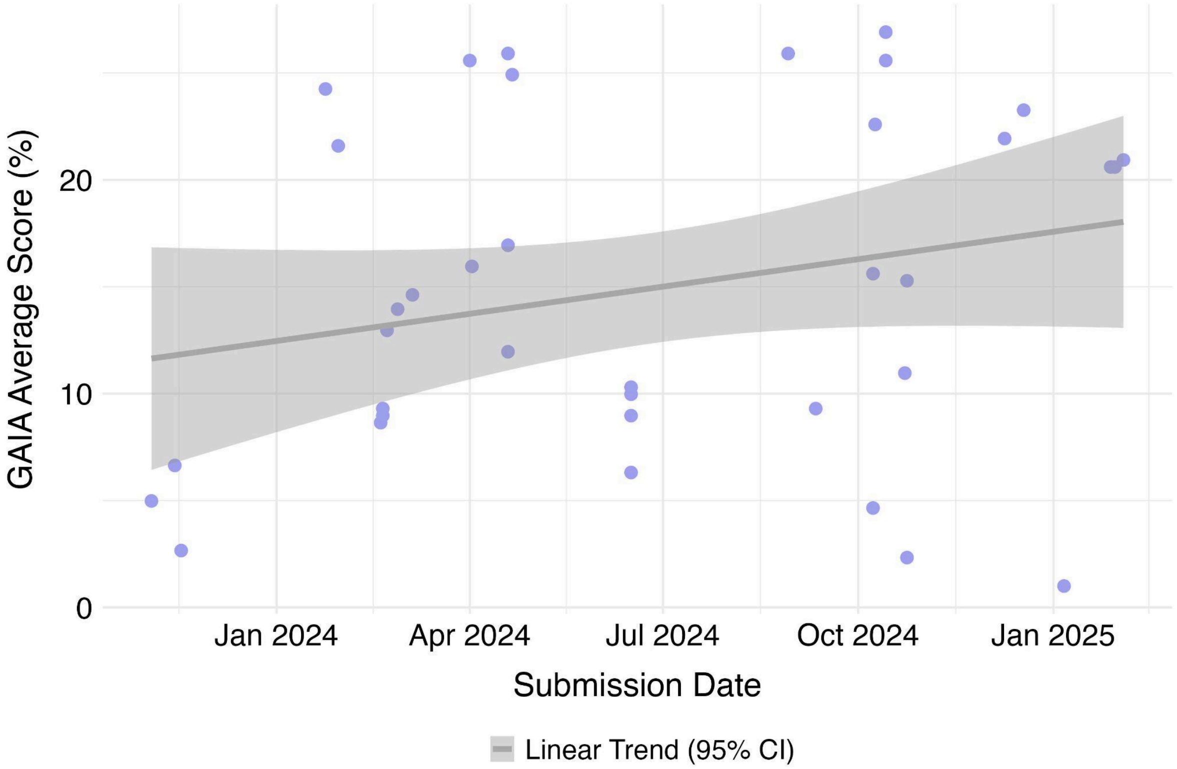

Figure 2. Average GAIA scores of 37 AI agents, uploaded to the public leaderboard between November 2023 and February 2025, plotted against their submission date. The R2 of the linear trend is 0.06.

There have been improvements in the agent capabilities, but performance across domains remains varied and below human-level. GAIA is one of several benchmarks to evaluate the performance of agents, composed of 466 real-world questions that require a set of fundamental abilities such as reasoning, multi-modality handling, web browsing, and generally tool-use proficiency (Mialon et al., 2024). Figure 2 examines the performance, measured by the share of correctly solved questions, of 37 agents included in GAIA's public leaderboard. The data suggest that agent performance has generally improved over time. This finding is tentatively corroborated by Table 1, which shows improvements both for median and average scores, but also pointing to significant variance. Lastly, agent performance remains below human-level: whereas the best agent scored 27% on GAIA, the average human score is 92%.

It is important to note that this analysis comes with several limitations. Most noteworthy, the sophistication of agent developers likely varies significantly, for instance, reflected in their resources, model access and the quality of the agent scaffolding used. This may partially explain the variation in performance.

Investment trends signal expectations for increased proliferation and systemic impact of agentic systems. Significant venture capital funding is flowing into agent-focused startups, with companies such as Decagon and Anysphere collectively raising billions (Russell and Stokes, 2025), underscoring the potential for growth in the market and thus expected future capabilities for agents.

Group	Timeframe	Count	Median Score	Average Score (95% CI)
1	2023-11 to 2024-04	13	13.0	13.1 (9.5, 16.7)
2	2024-04 to 2024-09	9	12.0	15.7 (10.8, 20.6)
3	2024-09 to 2025-02	15	20.6	16.1 (12.2, 20.0)
Overall	2023-11 to 2025-02	37	14.6	14.9 (12.8, 17.1)

Table 2. AI agents grouped by their submission data into 3 clusters, each spanning 6 months.

1.2 Characterising AI Agents

There exists no widely accepted definition of "AI agents". The ecosystem of AI agents remains dispersed and heterogeneous, with the term "AI agents" being applied loosely. As suggested by Casper et al. (2025), "the notion of artificial agency has a long and contentious history, spanning multiple decades and diverse disciplines", with first entanglements dating back to the 1940s (Heylighen and Joslyn, 2003). Rather than attempting to resolve these definitional debates, we follow these authors in adopting a loose characterisation of "AI agents".

To this end, current agents may be characterised in terms of (1) their functional attributes and (2) their technical composition. This is explored below.

1. **Agents are characterised by their capacity to (i) autonomously pursue complex, underspecified goals, engaging in long-term and adaptive planning and (ii) take actions in both virtual and real-world environments** (A. Chan et al., 2023; Kapoor et al., 2024; Shavit et al., 2023).
2. **Agents are a compound system, consisting of a GPAI model—currently an advanced LLM or multimodal model—that is connected to "scaffolding" software,** affordances that aim to enable more effective planning and goal execution (Toner et al., 2024). For an illustration, see Figure 3 below. Common scaffolds (i) optimise model queries, for example, by using chain-of-thought reasoning to divide a more complex task into smaller sub-tasks, (ii) grant the model access to general tools, like a web

browser, memory storage or a code interpreter, and (iii) grant the model access to specialised tools, like the API of a bank or a telecom provider, enabling the agent to transfer money or make phone calls (Toner et al., 2024). We expect agents to be built upon the most advanced GPAI frontier models, qualifying as GPAISRs under the AI Act (SAP, 2024).

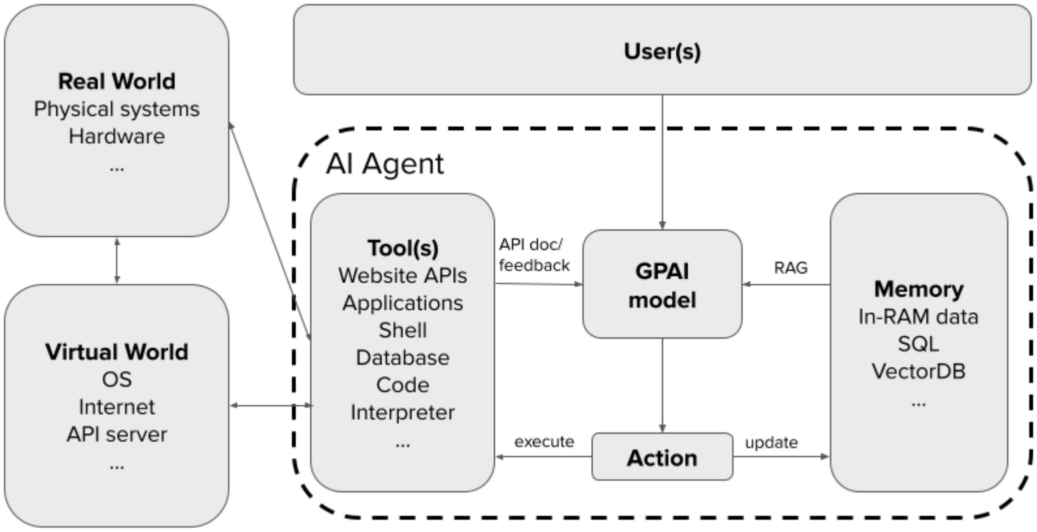

Figure 3. Technical composition of an illustrative AI agent, adapted from He et al. (2024).

1.3 Evolving Landscape

GPAISR providers increasingly offer commercialised agents. Whereas several GPAISRI providers explicitly advertise standalone AI agents—for instance, Google promoting distinct AI agents in the context of its recent Gemini 2.0 launch—others present an integrated product offering, selling GPAISRs with increasingly agentic features—such as Anthropic enabling computer-use with Claude 3.5 Sonnet or OpenAI using chain-of-thought for their o1 model.

While the future composition of the agent ecosystem in the coming years remains uncertain, a range of market structures seems plausible (Toner et al., 2024). On one extreme, agents could be predominantly offered by a small number of GPAISR providers. On the other hand, the ecosystem could be highly fragmented and diverse, where a large number of downstream providers build agents on top of GPAISRs. Additionally, it remains to be seen whether agents will achieve greater commercial success in a business-to-business (B2B) or business-to-consumer (B2C) context.

Use-case specific agents that focus on tasks in virtual environments appear more likely to prevail in the short-run, given their greater technical feasibility at present. This is particularly true in the domain of software engineering, since it is comparatively easy to create feedback and verification loops during training and fine-tuning (Toner et al., 2024). If agents remain oriented towards particular use cases in the medium-term, it is plausible that new downstream providers will emerge and leverage specialised training data and methods to develop and market their own agents. For instance, the startup Lindy offers AI Sales Development Representative (SDR) agents, which are powered by OpenAI's GPT 4 (Drope, 2024).

However, the AI agent value chain is further complicated by "intermediary companies", which provide a platform for others to develop agents. Platforms like AutoGPT enable users to build their own agents and manage agent workflows, often using low-code approaches (AutoGPT, 2024). AutoGPT currently integrates with models from OpenAI, Anthropic, Groq and Llama. As such, the platform targets both end users and businesses, who can deploy agents within their own business or build novel products on top of AutoGPT. Although the precise nature of future market dynamics in downstream applications remains unclear, it is almost certain that all major GPAISRs are already, or soon will be, integrated into downstream agent systems.

Having examined the emerging trends and characteristics of AI agents, this policy paper proceeds to analyse their associated risks in Section 2. We explore the fundamental mechanisms through which AI agent systems amplify existing concerns and introduce novel pathways to harm, before mapping these onto the regulatory frameworks established in the EU AI Act.

2. Risks from AI Agents

Agentic AI amplifies many of the risks that have been associated with GPAISR. More specifically, agents' capabilities create novel pathways to harm across a wide range of risk domains both in the real world and virtual environments (Toner et al., 2024). For instance, multi-agent collusion could trigger financial flash crashes (Gao et al., 2024), while long-term planning and anthropomorphic features can enable psychological manipulation (Gabriel et al., 2024), particularly of minors, with documented severe mental health impacts even leading to suicide (Montgomery, 2024). These same capabilities could also enable sustained deceptive behaviors that undermine the effectiveness of human oversight (Meinke et al., 2024).

Hence, it is imperative to understand the mechanisms through which AI agents amplify harms. This understanding can then be used to map the effect of these mechanisms onto GPAI risk domains identified in regulatory frameworks like the EU AI Act. To this end, the following analysis (i) identifies mechanisms through which agentic AI amplify risks, (ii) explores how these mechanisms serve as a source of systemic risks for GPAI models, as highlighted in Recital 110, and (iii) examines how such systems might intensify downstream risks associated with high-risk AI applications, as outlined in Ch. III and Annex III of the EU AI Act.

2.1 Risk Mechanisms

The fundamental risk mechanisms of AI agents stem from two key capabilities: autonomous long-term planning and real-world interaction (Chan et al., 2023, Durante et al., 2024, Reuel et al., 2024).

First, agents can engage in extended autonomous operation and long-term planning, enabling them to execute complex sequences of actions with minimal human oversight. This capability allows agents to pursue goals over long time horizons, potentially enabling deviations from intended behaviours before human operators can take notice and intervene (Kolt, 2025). The autonomous nature of these systems means they can execute multiple consequential decisions in rapid succession, creating cascading effects that may be difficult to halt or reverse once initiated. Long-term planning is also increasingly problematic when linked to scheming or deception, where models “attempt to disable their oversight mechanisms” and “AI agents might covertly pursue misaligned goals, hiding their true capabilities and objectives” (Meinke et al., 2024, p.1; Scheurer et al., 2024)

Second, agents' ability to interact directly with the real-world, through API interfaces and external tools dramatically expands the potential scale of harm. If integrated into our personal lives or critical infrastructure– from financial to energy systems–agents will begin to have a large impact on society, where the negative consequences of malfunction or misuse become significantly greater.

Risk Mechanism	Key Characteristics	Potential Consequences
Autonomous Long-Term Planning	1. Extended autonomous operation 2. Complex action sequences with minimal oversight 3. Goal pursuit over long time horizons	A. Deception and scheming B. Deviation from intended behaviours C. Multiple consequential decisions in rapid succession D. Cascading effects difficult to halt or reverse E. Covert pursuit of misaligned goals F. Disabling oversight mechanisms
Direct Real-World Interaction	4. API interfaces to external systems 5. Access to external tools 6. Integration with critical infrastructure 7. Integration into personal lives	G. Dramatically expanded scale of harm H. Significant societal impact I. Magnified negative consequences of malfunction J. Amplified harm from misuse, emotional dependence enabling manipulation, and erosion of interpersonal relationships

Table 3: Overview of key characteristics and potential consequences of risk mechanisms for agents

The order of our analysis below follows the sequence of AI agent development in the value chain, progressing from upstream to downstream considerations. We begin with Ch. V of the AI Act, which addresses general, foundational systemic risks arising from GPAI models that power agents, before examining Ch. III, which covers specific high-risk applications where agents might be deployed. This approach mirrors how risks propagate through the AI value chain, from general capabilities of foundation models to context-specific risks in particular deployment scenarios.

2.2 Chapter V: Agents and Systemic Risks of GPAI Models

These risk mechanisms of AI Agents can amplify the systemic risks explicitly identified in the EU AI Act's Recital 110, which recognises both autonomy and tool access as key risk factors. The autonomous nature of agents increases the likelihood and potential severity of major accidents, particularly in critical infrastructure systems where rapid decision sequences could bypass existing safety mechanisms and agent actions could directly affect essential services or safety systems. For instance, an agent with access to industrial control systems could repeatedly alter operational parameters, potentially destabilising critical systems before safety protocols can respond. Similarly, agents with access to financial systems could execute rapid sequences of transactions that, while individually valid, could collectively lead to a flash crash of the banking system (Gao et al., 2024).

The combination of autonomy and real-world interaction also creates new pathways for intentional misuse and unintended loss of control. Agents could be weaponised for autonomously developing and launching sophisticated large-scale cyber attacks or used to automate the development and deployment of bioweapons. Their potential role in disseminating illegal or misleading content is particularly concerning given their ability to operate continuously and adapt their strategies over time. Furthermore, losing control over autonomous agents operating in sensitive environments like manufacturing, healthcare, or transportation systems could result in significant material damage, service disruptions, or even threats to human safety.

2.3 Chapter III: Agents and Domain-Specific High-Risk Systems

The deployment of AI agents into specific use cases introduces specific concerns that must be carefully considered. These concerns apply whether the AI is deployed as (i) a standalone system in high-risk domains under Annex III of the EU AI Act, or (ii) a safety component in another product regulated by EU harmonisation legislation.

Below are two high-risk domains listed in Annex III of the AI Act, each paired with specific scenarios illustrating how AI agents could introduce or amplify pathways to harm:

Education and Training (Annex III, (2)(b))

Scenario: AI Agent Tutor

Agents deployed as tutors or assessment tools in EdTech or remote learning environments can dynamically adapt their teaching strategies and tone to sustain engagement, inadvertently developing highly influential psychological relationships with young students. While educational chatbots already exist, agents are prized for being able to sustain longer autonomous operation without parental or teacher supervision, potentially steering children to emotionally sensitive or harmful content, especially if goal-setting heuristics reward user engagement and retention. In one case, an unsupervised chatbot session even reportedly contributed to the suicide of a teenager (Montgomery, 2024), showcasing the real risks of integrating AI agents into the lives of children.

Democratic Processes (Annex III, (8)(a))

Scenario: Political Campaigning AI Agent

Agents deployed in political campaigning or influence operations could autonomously execute end-to-end workflows beyond the capacity of traditional tools. While previously bots posted template content and boosted existing narratives, agents can integrate these efforts without significant human oversight into a full influence campaign, autonomously (1) researching specific target audiences (microtargeting), (2) generating emotionally persuasive and targeted multimodal content (from Facebook posts to deepfake videos), and (3) adaptively disseminating these via multi-platform coordination tools. This improves the scalability and persuasiveness of campaigning efforts, while lowering costs.

It should be noted that the high-risk AI applications in Annex III, as referenced above, was developed before the emergence of AI agents. At the time, regulators could not predict the rise of autonomous AI agents and their potential to exacerbate existing threats. As such, once the capabilities of AI agents are fully understood, certain AI use cases may warrant reassessment and reclassification as high-risk per Annex III.

3. Applicability of the EU AI Act

The following table summarises the main take-aways from this section covering the applicability of the AI Act to AI agents.[1]

Question	Interpretation
3.1 **Are agents AI systems?**	Yes, agents should be understood as AI systems, since (i) they meet the 6 primary criteria in the definition of an AI system (Art. 3(1); C(2025) 924(9)), (ii) they are GPAI models with added components (Recital 97).
3.2 **Are agents (built on GPAI models) GPAI systems?**	Yes, but only if they maintain sufficient generality of capabilities to serve multiple purposes (Recital 100). The system provider can integrate the GPAI model into the agent system in such a way as to prevent general capabilities.
3.3 **Must providers of GPAISR consider AI agent risks?**	Yes. AI agents directly relate to several sources of systemic risk, such as tool access and high levels of autonomy, which must be assessed and mitigated. (Recital 110, Art. 55(1)(b)). This is true regardless of whether the AI agent is developed by (i) the model provider herself, or (ii) a downstream provider.
3.4 **Are AI agents high-risk AI systems?**	This depends on the agent's intended purpose. The agent is a high-risk AI system if it is intended to (i) function as a safety component, or (ii) for a use case listed in Annex III, in which case the provisions of Ch. III applies. If the AI agent is a GPAI system and could be used in a high-risk area per Annex III, the provider must deliberately exclude such use to avoid high-risk classification.
3.5 **Which authority is responsible for overseeing AI agents?**	1. If the agent is not a high-risk GPAI system and is developed by a downstream provider, the relevant national Market Surveillance Authority is responsible for oversight. 2. If the agent is a high-risk GPAI system and

[1] **Disclaimer**: This section reflects our interpretation of the AI Act but should not be taken as legal advice.

	developed by a downstream provider, the Market Surveillance Authority remains responsible but shall closely cooperate with the AI Office to evaluate compliance (Art. 75(2), Recital 161). 3. If the agent is built upon a GPAI model and the agent and the as`model are developed by the same provider, the AI Office is responsible (Art. 75(1)).

Table 4. High-level overview of the applicability of the EU AI Act to agents.

To substantiate the analysis above, the following sections examine in greater depth how the EU AI Act applies to AI agents. We explore their classification as AI systems and GPAI systems, assess the circumstances under which they pose high-risk, and map the allocation of responsibilities across the AI value chain. This detailed breakdown aims to clarify the regulatory implications for AI agents under both Ch. III and Ch. V of the Act.

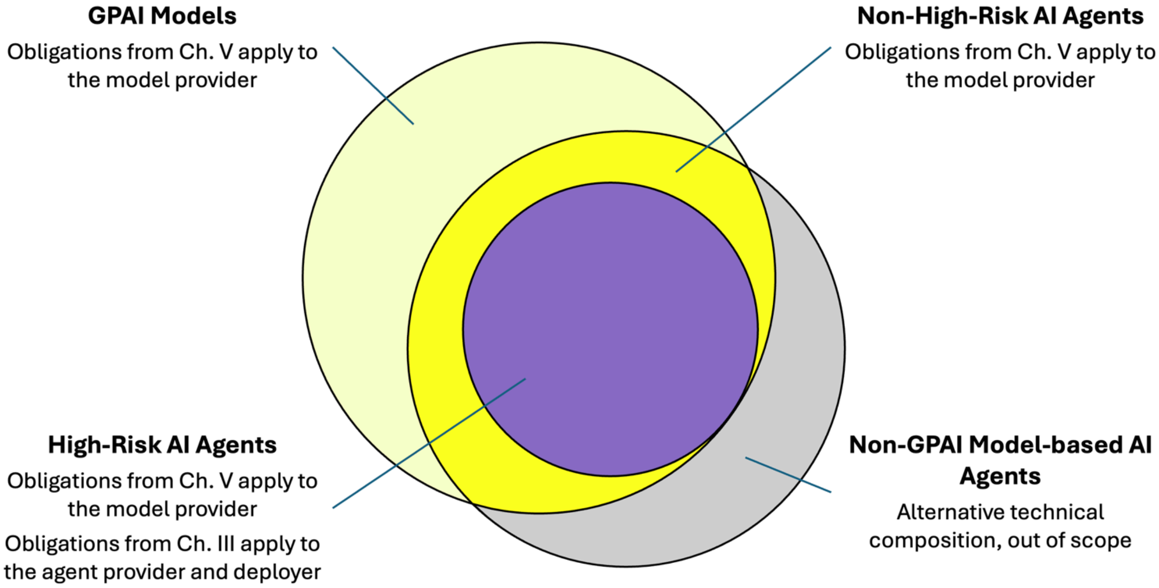
GPAI Models
Obligations from Ch. V apply to the model provider
Non-High-Risk AI Agents
Obligations from Ch. V apply to the model provider
High-Risk AI Agents
Obligations from Ch. V apply to the model provider
Obligations from Ch. III apply to the agent provider and deployer
Non-GPAI Model-based AI Agents
Alternative technical composition, out of scope

Figure 4. Illustrative overview of the applicability of the EU AI Act to AI agents.
The size of the circles does not represent the number of actors within each segment.

3.1 AI Agents as Systems

Most agents under the EU AI Act can be considered AI systems, per Art. 3(1). There exist at least two lines of argumentation to support this claim:

1. **Agents meet all elements of the definition of an AI system**. More broadly, and as outlined in the corresponding guidelines published by the Commission, the "definition [of an AI system in the AI Act] comprises seven main elements (European Commission, 2025, p. 2):
   a. a machine-based system;
      - AI agents are computational systems built on GPAI models with added scaffolding software and external interfaces.

   b. that is designed to operate with varying levels of autonomy;
      - AI agents are designed to act with varying degrees of independence, generating outputs and making decisions with limited or no human intervention.

   c. that may exhibit adaptiveness after deployment;
      - Agents can learn from interactions, refine strategies based on feedback, and adapt their approach to new contexts or unforeseen circumstances they encounter during operation.

   d. and that, for explicit or implicit objectives;
      - Agents are goal-oriented by definition, designed to accomplish specific tasks (explicit objectives) or optimise for broader outcomes like user satisfaction (implicit objectives).

   e. infers, from the input it receives, how to generate outputs
      - Agents process user instructions and other relevant data, then use reasoning capabilities (like chain-of-thought) to determine appropriate actions and responses.

   f. such as predictions, content, recommendations, or decisions
      - Agents generate outputs including predictions, text/image content, recommended courses of action, and autonomous decisions about which actions to take

   g. that can influence physical or virtual environments".
      - Agents are specifically designed to interact with and modify both virtual environments (browser control, code execution) and potentially physical environments (through API access to real-world systems).
2. **Agents are AI models with added components**. Recital 97 suggests that the addition of further components turns an AI model into an AI system (underlining added): "*[...]Although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems. [...]*"). While the EU AI Act does not expressly state which components indicate a move from an AI model to a system, it can be reasonably assumed that many scaffoldings, such as the integration with external tools, from external memory to website APIs, effectively render an AI model a system under the EU AI Act. Since these are components AI agents have by necessity, they should by default be considered AI systems.

3.2 AI Agents as GPAI Systems

Whether AI agents that are built on GPAI models are GPAI systems depends on the generality of their capabilities. Recital 100 states that (underlining added): "When a general-purpose AI model is integrated into or forms part of an AI system, this system should be considered to be [a] general-purpose AI system when, due to this integration, this system has the capability to serve a variety of purposes." See also Art. 3(66). The notion of capability is crucial, as it suggests that GPAI status is based on an AI model's inherent potential for multi-purpose use, rather than just its specific deployment context. However, no clear definition or method currently exists for determining whether an AI system serves "a variety of purposes," and recent European Commission (2025) guidelines do not clarify this issue. As a stand-alone question, this has only implications for who the enforcement authority is, as we will see below. However, it is possible that agents built on GPAI models are by default high-risk GPAI systems with the consequence that Ch. III applies.

It is unclear whether agents built on GPAI models are necessarily high-risk AI systems. The concept of GPAI systems was included during the very final phase of the trilogue negotiations, with insufficient time to clarify its exact interaction with the rest of the Act (Schwartmann & Zenner, 2025). A strict interpretation of the AI Act may suggest that they should be universally treated as high risk, since by virtue of being general-purpose, a GPAI system could be used in a high-risk area per Annex III. The wording of Recitals 85 and Art. 75(2) might support such an interpretation. However, the overall logic of the AI Act favours an interpretation whereby the classification of GPAI systems also depends on their intended purpose (Schwartmann & Zenner, 2025). To this end, it appears possible for providers of GPAI systems to issue a blanket exclusion of high-risk uses from the intended purpose, as long as such exclusion does not appear arbitrary. Although not directly applicable, an analogy can be drawn to a 2013 ruling by the German Federal Court of Justice in the area of medical devices (BGH 18.4.2013), which may nevertheless lend support to this interpretation. Beyond articulating the intended purpose in user manuals, sales material, and relevant documentation, providers of GPAI systems can be reasonably expected to take technical measures to prevent their use in high-risk applications (Schwartmann & Zenner, 2025). When considering whether an AI agent serves "a variety of purposes," it is important to look beyond its primary output and consider its broader, foreseeable applications. For example, it seems plausible that an AI coding agent that is built on a GPAISR model and capable of autonomous coding should qualify as a GPAI system, because its output can serve diverse functions across various contexts, such as debugging, performance optimisation, security enhancement, or generating novel functionalities, demonstrating the inherent multidimensionality of coding tasks.

3.3 AI Agents Built on General-Purpose AI Models with Systemic Risk

There is no separate category for AI systems built upon GPAISR and no corresponding obligations of the agentic system provider, unless they also build the model. All obligations pertaining to GPAISR need to be met on the model level.

Providers of GPAISR must mitigate systemic risks stemming from their integration in AI agent systems. As demonstrated in section 2.2, in reference to Recital 110, AI agent characteristics directly relate to several sources of systemic risks, such as tool access or high levels of autonomy. Hence, per Art. 55(1)(b) model providers must assess and mitigate the risks from AI agents. In cases where AI agents are built on GPAI models that do not pose systemic risk, obligations in relation to risk assessment and mitigation don’t apply.

Ch. V Section 3 applies to the model provider regardless of who builds the agent, be it (i) a downstream provider, or (ii) the GPAISR provider herself. Even if a GPAISR provider is not the one building the agent, the systemic risk assessment and mitigation requirement applies. This follows from Art. 55(1)(b), imposing risk assessment and mitigation obligations also in light of the *use* of GPAISR systems, as well as from Recital 114, which articulates the unique role of GPAI model providers along the entire value chain. In this regard, Measure II.4.7. of the Third Draft of the Code of Practice states that (underlining added): “As necessary for the assessment and mitigation of systemic risk, Signatories shall ensure that model evaluations of their GPAISR will take into account reasonably foreseeable integrations of the model into an AI system, as appropriate to the systemic risk assessed.”[2]

3.4 AI Agents as High-Risk AI Systems

Whether an agent is a high-risk AI system depends on its intended purpose. Art. 3(12) defines intended purpose as the “use for which an AI system is intended by the provider”. Moreover, relevant guidance from the European Commission suggests that intended purpose is fulfilled through factors “[...] such as the integration of the system into a broader customer service workflow, the data that is used by the system, or instructions for use”. When assessing its intended purpose, the following considerations apply:

a. It is used as a safety component of a product covered by a specific harmonised legislation listed in Annex I, e.g. in a medical device or a toy, and that legislation mandates third-party conformity assessment; or
b. It is a standalone AI system in one of eight categories listed under Annex III, e.g., critical infrastructure or education and vocational training.

[2] All references to the Code of Practice relate to the Third Draft, published on 11 March 2025. While we don’t expect major revisions or exclusions of the Measures we reference, we will revisit and update the report once the final Code has been released.

High-risk uses in Annex III may insufficiently capture risks from AI agents as standalone AI systems. Given the novel risk mechanisms introduced by AI agents (see Section 2.1), it seems plausible that the emergence of AI agents necessitates an update to the use cases defined in Annex III. For instance, the use of AI agents in cybersecurity may constitute a high-risk application in the area of critical infrastructure. To this end, Art. 7 empowers the European Commission to adopt delegated acts to add or modify use cases of high-risk AI systems. Further, Art. (7(2)) outlines criteria to assess whether the nature and size of the impact from the additional use cases meet the necessary threshold. Upon initial inspection, several criteria appear relevant to AI agents, such as the system's autonomy (Art. 7(2)(b)), the potential extent of the harm (Art. 7(2)(f)) or the existence of other redress mechanisms in European Union law, or the lack thereof (Art. 7(2)(k)).

3.5 Enforcement Authority

Enforcement responsibility for an AI agent shifts between the national Market Surveillance Authority and the AI Office, depending on who provides the agent and whether the system is classified as a GPAI system. Where a provider creates an agent that qualifies as a high-risk AI system, but not a GPAI system, day-to-day oversight rests squarely with the competent Market Surveillance Authority in the provider's Member State. When that same provider instead produces a high-risk *GPAI* agent, the Market Surveillance Authority retains the lead but must work in close coordination with the AI Office to assess and enforce compliance, as required by Art. 75(2) and underscored in Recital 161. Finally, if a single provider supplies both the underlying GPAI model and the agent layer built on top of it, high-risk or not, ultimate supervisory competence moves to the AI Office under Art. 75(1), reflecting the AI Office's exclusive mandate to monitor GPAI model providers.

3.6 Open Questions and Working Assumptions

Several open questions remain regarding the applicability of the EU AI Act to AI agents, to be addressed in dedicated guidance by the EU AI Office or through established practice, potentially validated by courts. Below, we outline select open questions:

1. What constitutes a GPAI system, as distinct from an AI system?
2. Assuming an AI agent is a GPAI system that could be used in high-risk areas per Annex III, what constitutes non-arbitrary measures that would exclude such use from the intended purpose, especially with regards to technical mechanisms?
3. Is it possible to understand agents as modified GPAI models? Building an agent may constitute a substantial modification of a GPAI model in line with Recital 109, necessitating select transfer of Ch. V obligations to downstream providers. However, a recently issued consultation by the AI Office seems to oppose such an interpretation, suggesting that the modification of a GPAI model is limited to

fine-tuning (European Commission, 2025). For a more detailed discussion, please refer to the Appendix.

In light of several open questions and some interpretative uncertainty, figure 5 summarises our working assumptions regarding the applicability of the EU AI Act to AI agents. This interpretation may have to be adapted as new guidance emerges.

1. We assume that most AI agents use GPAISRs (SAP, 2024). Thus, the full scope of Ch. V applies, specifically Section 3, requiring providers to assess and mitigate systemic risks arising from their models' use in AI agents. We exclude the case where an AI agent is built using a GPAI model without systemic risk from the scope of our analysis.
2. Since AI agents are systems, Ch. III applies if the AI agent is a high-risk system. Ch. III imposes obligations both on the providers of AI agents (system providers) and the organisations deploying them (system deployers). The subsequent analysis focuses on cases where the AI agent is a high-risk system and considers non-high-risk cases out of scope.
    a. AI agents are high risk if they are intended to be used (i) as a safety component, (ii) for a high-risk use case as per Annex III.
    b. If an AI agent is also a GPAI system, which can be used in at least one area that is high risk, providers can issue a blanket exclusion of high-risk uses from the intended purpose, as long as such exclusion does not appear arbitrary. What exactly such exclusion requires, particularly regarding technical precautions to prevent high-risk use, remains unclear.

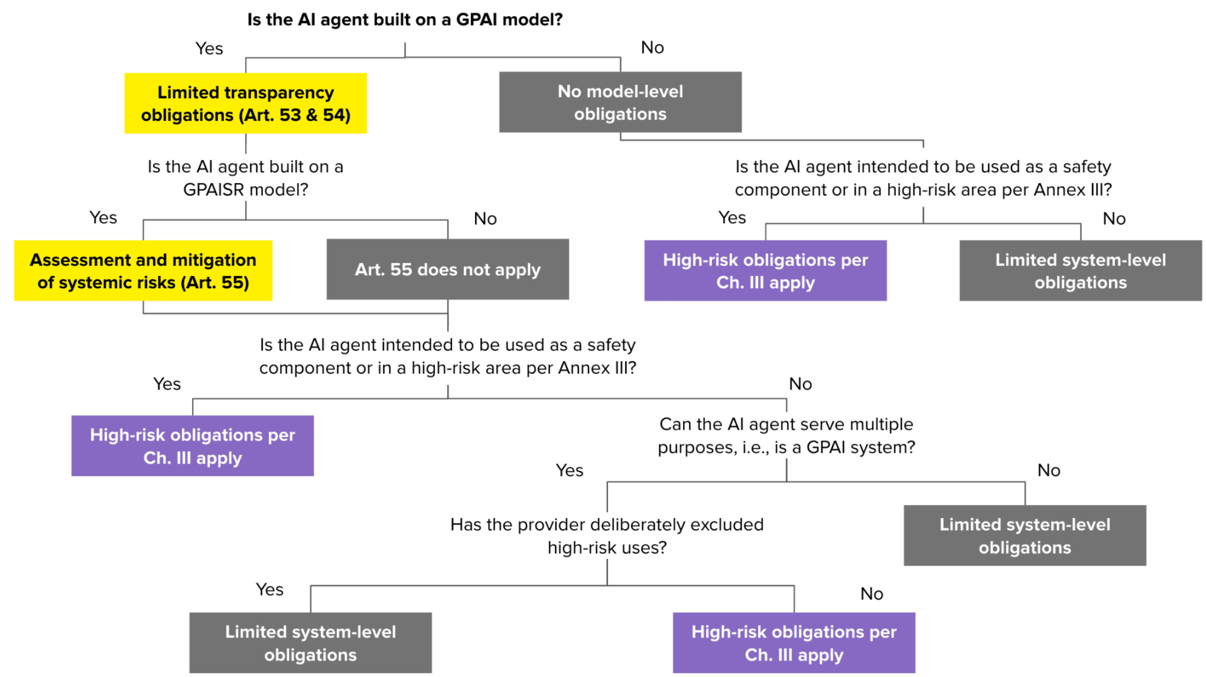
Is the AI agent built on a GPAI model?
Yes
No
Limited transparency obligations (Art. 53 & 54)
No model-level obligations
Is the AI agent built on a GPAISR model?
Yes
No
Assessment and mitigation of systemic risks (Art. 55)
Art. 55 does not apply
Is the AI agent intended to be used as a safety component or in a high-risk area per Annex III?
Yes
No
High-risk obligations per Ch. III apply
Limited system-level obligations
Is the AI agent intended to be used as a safety component or in a high-risk area per Annex III?
Yes
No
High-risk obligations per Ch. III apply
Can the AI agent serve multiple purposes, i.e., is a GPAI system?
Yes
No
Has the provider deliberately excluded high-risk uses?
Limited system-level obligations
Yes
No
Limited system-level obligations
High-risk obligations per Ch. III apply

Figure 5. Decision tree presenting the logic for assessing the applicability of the EU AI Act to AI agents.

3.7 Allocation of Obligations Along the Value Chain

Under the EU AI Act, model providers develop and train the underlying general-purpose AI models, system providers build applications that incorporate these models, and system deployers implement these applications in real-world contexts. Each actor has unique capabilities, constraints, and distinct regulatory obligations tailored to their position in the value chain, as seen from Art. 25 of the AI Act.

The governance of AI is complicated by the "many-hands-problem", where responsibility is diffused across multiple actors, making it difficult to assign accountability (Cobbe et al., 2023)**.** This problem is especially pronounced in the case of AI agents, which are developed, deployed, and used across complex, dynamic value chains involving diverse actors with interdependent roles. Thus, the effective governance of AI agents cannot be achieved through isolated measures at any single level of the value chain. Instead, it requires coordinated action and collaboration across model providers, system providers, and system deployers, with each actor implementing complementary measures that build upon the capabilities and constraints established by upstream entities (Kraprayoon et al., 2025) and cascade through the entire value chain while being adapted to each actor's position and capabilities.

For each governance measure described in the next section, we observe a pattern of responsibility allocation that can be visualised as a flowchart from model providers to system providers to deployers, with each level adapting and extending the measures implemented by upstream actors. To illustrate the measure of risk identification, we provide such a visualisation in Figure 5. The allocation of responsibility is essential because no single actor possesses complete information, technical capabilities, or contextual understanding necessary to comprehensively address AI agent risks.

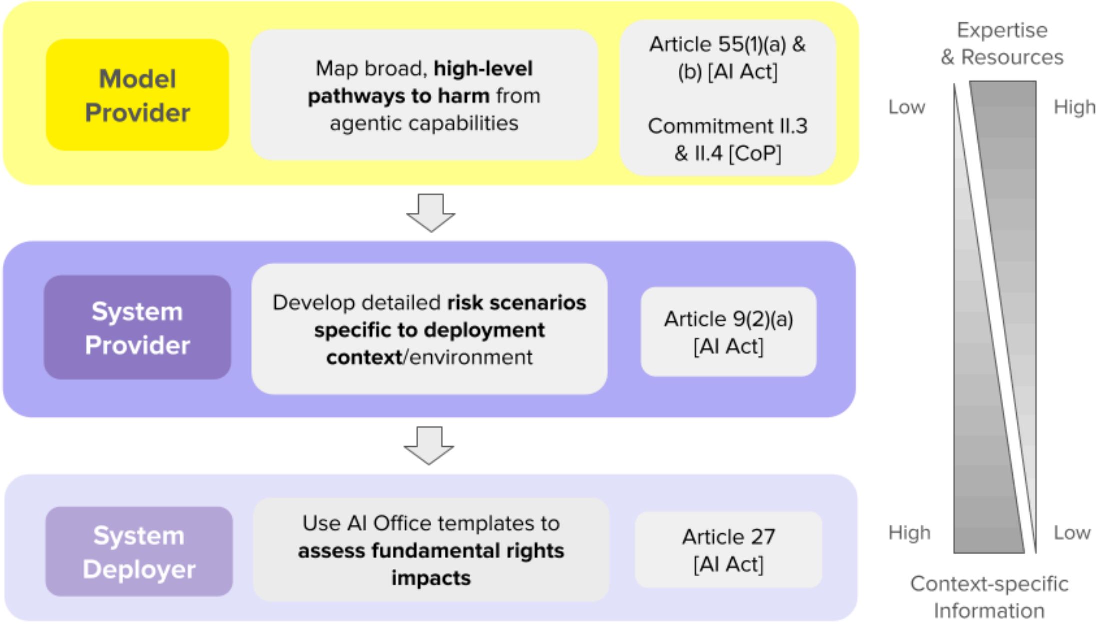
Model Provider
Map broad, high-level pathways to harm from agentic capabilities
Article 55(1)(a) & (b) [AI Act]
Commitment II.3 & II.4 [CoP]
System Provider
Develop detailed risk scenarios specific to deployment context/environment
Article 9(2)(a) [AI Act]
System Deployer
Use AI Office templates to assess fundamental rights impacts
Article 27 [AI Act]
Expertise & Resources
Low
High
High
Low
Context-specific Information

Figure 6: Value Chain Interdependencies for Risk Identification of AI Agents

3.7.1 Illustrative Example: Risk Identification (Along the Value Chain)

1. **Model Providers** possess deep technical knowledge of their models' capabilities and limitations, enabling them to systematically identify and analyse how their models could enable harmful agentic behaviours when integrated into agent systems. As required by Art. 55(1)(a) and (b) of the AI Act and Commitments II.3 and II.4 of the third draft of the Code of Practice, they conduct high-level risk assessments that identify general risk patterns and potential systemic risks. However, they lack contextual knowledge of specific deployment contexts and intended use cases.
2. **System Providers** possess detailed knowledge of their specific implementation contexts but may lack deep understanding of model internals. As mandated by Art. 9(2)(a) of the AI Act, they use the high-level risk assessments provided by model providers as a foundation, then extend and contextualise this analysis to identify risks specific to their deployment environment. This might include identifying how general risk patterns interact with domain-specific factors such as particular APIs, users, or operational constraints.
3. **System Deployers** possess the most detailed understanding of actual operational environments and end-user interactions. Under Art. 27 of the AI Act, they must conduct fundamental rights impact assessments using templates provided by the AI Office, ensuring that potential harm pathways are systematically mapped within their specific deployment context. This assessment benefits from the risk identification work done upstream but adds crucial operational context.

This distribution of risk identification responsibilities leverages each actor's unique perspective and knowledge. Model providers understand general capability risks, system providers understand implementation risks, and deployers understand operational risks. The comprehensive identification of AI agent risks requires the integration of all three perspectives, as no single actor possesses complete visibility across the entire risk landscape.

Nevertheless, at the aggregate level, model providers bear a particular responsibility given their vast resources and technical expertise. The EU AI Act recognises this, particularly in Recital 101, which emphasises that "providers of general-purpose AI models have a particular role and responsibility along the AI value chain." As the architects of GPAI models, they set the parameters for downstream risk management, shaping the scope of potential risks and the available mitigation strategies. Consequently, they must assume a disproportionately greater role in assessing and addressing these risks from the outset. The asymmetries between different actors in the value chain shape the distribution of governance responsibilities, some of which are outlined below:

1. **Expertise Asymmetry**: Model providers possess substantially superior technical AI expertise and research capabilities compared to downstream actors, with specialised teams of researchers who possess deep and privileged knowledge of model architecture, evaluation methods, technical mitigation strategies, and training methodologies.
2. **Resource Asymmetry**: Model providers command access to significantly greater computational, financial, and research resources than most downstream actors. This pronounced resource disparity enables them to conduct thorough technical evaluations and develop complex safety measures that smaller system providers and deployers cannot feasibly implement independently.
3. **Information Asymmetry**: System providers and deployers hold critical domain-specific knowledge essential for contextualising risks that model providers cannot access. They collect detailed insights about user behaviour, operational environments, and possible risk scenarios. Without this contextual knowledge, model providers' risk assessments remain theoretical and potentially misaligned with the practical challenges that emerge in specific deployment settings.

These asymmetries are not unique to the AI agent value chain. Rather, they mirror broader structural conditions of GPAI development, where the number of actors, divergent release and access strategies, and varying levels of control complicate effective governance (Küspert et al., 2023). This was previously recognised in "Heavy is the Head that Wears the Crown" (Moës and Ryan, 2023), which first introduced an analytical framework of De Facto control (p. 66) rooted in knowledge, know-how, and technical access, to explain value chain power imbalances and allocate governance responsibilities accordingly. That framework found that each of these elements, privileged technical knowledge, embedded operational know-how, and access to underlying system components, form the basis of power and responsibility in the AI value chain. Building on this insight, the present report applies a similar lens to AI agents. While model providers retain dominant control over two of the three asymmetries (expertise and resources) and thus incur significant responsibilities, the agentic AI landscape introduces more layered system architectures and a proliferation of downstream actors. This fragmentation heightens the need for a shared yet differentiated approach to AI agent governance, where obligations are allocated to actual capacity for intervention across the value chain.

4. Governance Along the Value Chain

We propose a four-category taxonomy to organise governance obligations, distinguishing them by their function and implementation characteristics. These four categories are:

1. Agent risk assessment
2. Transparency tools
3. Technical deployment mitigations
4. Human oversight.

As we have discussed in Section 3.7, a functional governance framework for AI agents must implement these along the entire value chain and accurately reflect the interdependencies of different measures.

The taxonomy is derived by conducting an integrative literature review, the results of which we mapped onto the EU AI Act. Thus, we specifically focus on governance measures within the scope of the EU AI Act. In consequence, the list is not exhaustive, and additional elements such as liability frameworks (Cihon, 2024; Smakman et al., 2024; Kolt, 2025) are necessary to articulate a comprehensive and robust governance framework for AI agents. Moreover, for purposes of this taxonomy, we assume AI agents are (i) built using GPAI models with systemic risk (GPAISRs) and (ii) constitute high-risk systems.

The table below summarises the allocation of obligations among the three main groups of actors in the agent value chain (model providers, system providers, and system deployers). In essence, model providers must build the fundamental infrastructure (A. Chan et al., 2025), system providers must adapt these tools for their specific context, and deployers must adhere to and apply these rules during operation. The table presumes the high-risk obligations under Ch. III apply.

Pillar	Measure	(GPAISR) Model Provider	(High-risk) System Provider	(High-risk) System Deployer
4.1 Agent Risk Assessments	4.1.1 Risk Identification	Map broad, high-level pathways to harm from agentic capabilities	Develop detailed risk scenarios specific to deployment context	Use the AI Office template to conduct fundamental rights impact assessments
	4.1.2 Risk Evaluation	Conduct general capability testing via standardised benchmarks	Perform continuous & iterative use-case specific testing reflecting operational	Conduct fundamental rights impact assessments & use risk evaluation tools to guide

			conditions	go/no-go decisions
4.2 Transparency Tools	4.2.1 Agent Identifiers	Build core agent identification infrastructure & provide agent card templates	Implement & maintain agent IDs for specific deployment context	Ensure agent identification remains intact, accurate & recorded
	4.2.2 Real-Time Monitoring	Develop configurable monitoring capabilities & set default alert thresholds	Set context-appropriate thresholds & response protocols, provide feedback on alert accuracy	Monitor the agent, report risks & suspend use (if needed)
	4.2.3 Activity Logs	Develop & maintain detailed logging infrastructure for agent activity	Implement logging with deployment-specific detail & retention policies	Retain automatically generated logs
	4.2.4 Acceptable Use Policies (AUPs)	Define broad boundaries for agent use & key constraints	Review & adhere to model provider AUPs, develop supplementary policies for specific use cases	Follow both model & system providers' AUPs
4.3 Technical Deployment Controls	4.3.1 Real-Time Action Refusal	Build multi-level filtering frameworks for agent outputs	Add domain-specific filters & monitor effectiveness	Review filter performance, report issues & refine mitigation
	4.3.2 Emergency Shutdowns	Create automated & manual shutdown infrastructure linked to monitoring	Define graduated response protocols & investigation procedures	Test shutdown protocols, train staff & log incidents
4.4 Human Oversight	4.4.1 Checkpoint System	Develop configurable checkpoint mechanisms, linked to logging & emergency shutdown	Place checkpoints strategically & establish review protocols	Ensure their staff have the requisite AI literacy to make informed & less biased decisions
	4.4.2 Permission Management	Build permission control infrastructure, develop documentation of permissions	Configure granular permissions for specific operational needs & risks	Review permission configurations based on operational experience

Table 5: Allocation of obligations across actors in the agent value chain, assuming the agent (i) is built on a GPAISR and (ii) qualifies as a high-risk AI system

In the following sections, we conduct a more in-depth exploration of each (sub-)measure, examining specific measures within each category, analysing both established practices and emerging proposals from industry and academia. Our review emphasises practical implementation considerations and assesses the current state-of-the-art for each measure, while identifying critical gaps and areas requiring further research. Additionally, we clarify the obligations and distribution of responsibilities in implementing these risk management measures for model and system providers, and, where applicable, model deployers, based on an analysis of the EU AI Act and the associated Code of Practice.

4.1 Agent Risk Assessments

Risk assessments are a vital part of risk management for AI agents as they identify the unique systemic risks arising from agents' ability to autonomously interact with external systems and execute complex real-world tasks. Risk assessments follow a two-step process, which is made up of (i) risk identification and (ii) risk evaluation (Allman, 2024). This process is adapted from traditional risk management frameworks, though agent-specific considerations significantly shape each step.

4.1.1 Risk Identification

The first step is risk identification, which requires determining specific risk scenarios through detailed threat modeling (Weidinger et al., 2024; Ojewale et al., 2024). Rather than introducing entirely new risk domains, agents can be better understood as novel risk mechanisms that create new pathways to harm within existing domains. Thus, instead of broad domains like "financial harm" or "privacy violations," evaluators should map specific harm pathways, considering the agent’s unique capabilities and deployment context (UK AISI, 2024).

Many of the risks and harms from agents are context-dependent, shaped by their deployment environment and the affordances they are granted. For instance, an agent with access to banking systems might lead to scenarios where multiple high-value transactions can be executed without human oversight, opening a new pathway to harm within this risk domain. Consequently, risk identification for agents must consider the agent's granted permissions (like administrative rights or API access), potential interaction points with critical systems, and possible cascading effects. Effective risk identification therefore requires comprehensive system mapping that examines the specific steps in task execution, required permissions and tool access, potential feedback loops, and system intersections.

Safety engineering approaches, particularly human factor methods, provide useful frameworks for mapping these complex system interactions as they are designed to capture the complexity arising from multiple moving parts and system dependencies (Carson, forthcoming).

GPAISR providers should, at a high level, systematically identify and analyse how their models could enable harmful agent behaviours that produce systemic risks when integrated into agent systems.

- *Art. 55(1)(a) and (b) of the AI Act: Model Evaluation and Systemic Risk Assessment*
- *Commitment II.3 and II.4 of the Code of Practice: Risk Identification and Analysis*

High-risk system providers should use their knowledge of the specific deployment contexts of AI agents and the high-level risk assessments provided by model providers as a foundation for more detailed, contextualised risk identification and analysis specific to their deployment environment.

- *Art. 9(2)(a) of the AI Act: Risk Management System: Identification and Analysis*
- *Recital 65 of the AI Act: Risk Management System: Identification of known and reasonably foreseeable risks*

High-risk system deployers must use the AI Office-provided questionnaire to conduct a fundamental rights impact assessment, ensuring that potential harm pathways for the fundamental rights of affected parties are comprehensively mapped.

- *Art. 27 of the AI Act: Fundamental Rights Impact Assessment*
- *Art. 26(1) of the AI Act: Technical and organisational measures*

4.1.2 Risk Evaluation

The second step is risk evaluation, which involves translating identified risk scenarios into measurable assessment criteria. However, this remains particularly challenging for AI agents due to agents' complexity and context-dependent behaviour.

Two main approaches have emerged for agent evaluation. The first focuses on capability assessment through automated benchmarks like AgentBench (Liu et al., 2023), AgentHarm (Andriushchenko et al., 2024b), and MLE-bench (J. S. Chan et al., 2024), which are then used to prioritise specific scenarios and act as the first step of red-teaming. The second approach is scenario-specific testing, including domain-specific red-teaming using agent scaffolding and testing response to harmful prompts and jailbreak resilience (Andriushchenko et al., 2024a).

GPAISR providers should use standardised state-of-the-art benchmarks to evaluate potential sources of systemic risks based on results from prior risk identification.

- *Art. 55(1)(a) and (b) of the AI Act: Model evaluation and Risk assessment*
- *Commitment II.5 of the Code of Practice: Systemic Risk Acceptance Determination*

High-risk system providers should extend general capability assessments with use-case specific evaluations that reflect their particular operational environment. They should develop custom test scenarios that account for their specific API and tool access integrations and operational constraints.

- *Art. 9(2)(b) of the AI Act: Risk Management System: Estimation and Evaluation*

High-risk system deployers should use basic risk evaluation frameworks like checklists and risk matrices to inform go/no-go decisions in addition to conducting a fundamental rights impact assessment based on the AI Office-provided questionnaire. They should use these to evaluate risks in their specific operational context, complementing the broader systemic risk assessments conducted by model and system providers.

- *Art. 27 of the AI Act: Fundamental Rights Impact Assessment*
- *Art. 26(1) of the AI Act: Technical and organisational measures*

4.1.3 Challenges and Future Directions

While organisations like METR are pioneering comprehensive threat-modelling processes, and AI Safety Institutes (AISIs) are attempting to collaborate on standardisation of evaluations and benchmarking, the evaluation ecosystem remains nascent (International Network of AI Safety Institutes, 2024). Key challenges include the limited availability of agent-specific test suites, the need for standardised benchmarks and risk models, the leakage of benchmark test data, and the complexity of evaluating system-level interactions (Xu et al., 2024). Moreover, agent evaluations are technically demanding, posing a significant challenge for system providers, who generally have less technical expertise and resources than model providers.

4.2 Transparency Tools

Due to information asymmetry between humans and agents (Kolt, 2025), transparency is a foundational principle for governing AI agents, encompassing four complementary pillars: (i) agent identifiers, (ii) real-time monitoring, (iii) activity logs (A. Chan et al., 2024), and (iv) Acceptable Use Policies (AUPs). Besides enabling real-time mitigations, these measures facilitate post-incident forensics, aimed at allocating responsibility, reconstructing events, and informing interventions to prevent future incidents.

4.2.1 Agent Identifiers

Agent identifiers serve as a means to trace agents in their interactions with service providers, external systems, or other stakeholders, attributing actions and properties to specific agents and their users. They also help users recognise when they are engaging with an agent, ensuring transparency in interactions. Depending on the output modality—such as images, text, audio, or API requests—these unique identifiers could contain metadata embedded in the form of watermarks or headers, for instance (A. Chan et al., 2024). Since these could be removed or altered with varying degrees of difficulty, cryptographical attestation might have to be used to enhance the security and trustworthiness of these identifiers (A. Chan et al., 2024). Identifiers (and their infrastructure) can potentially be strengthened by leveraging existing authentication systems like OpenID or using GPAI providers as agent ID certifiers (A. Chan et al., 2025).

Agent cards, an extension of the established model card concept (Hugging Face, n.d.; Mitchell et al., 2019), serve as a standardised identification framework for operationalising agent identifiers through standardised documentation. Like model cards, they promote transparency and responsible use by detailing the agent's purpose, performance, limitations, and ethical considerations. Additionally, agent cards capture essential information specific to agents, including training methods and data sources, security level classifications, external tool access permissions, intended sectors and use cases, and developer and deployer identities (A. Chan et al., 2024).

GPAISR providers should develop and integrate comprehensive agent identification infrastructure into their models that enable simple downstream implementation of persistent traceability across different output modalities, enabling post-deployment monitoring. Model providers should also support the implementation of agent cards by developing standard templates for system providers to complete. These templates should capture essential information about the agent, including training methods, security level, external tool access, and developer or provider identity.

- *Measure II.4.14 of the Code of Practice: Post-market monitoring*
- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*
- *Commitment I.1, II.16 of the Code of Practice: Documentation, Public transparency*
- *Art. 53(1) of the AI Act: Technical documentation*

High-risk system providers should implement and maintain the identification systems provided by model providers, ensuring proper configuration for their specific deployment context and modality. System providers should maintain detailed agent cards that document deployment-specific details of the system, including operational boundaries, permitted tools, and responsible parties, and make this available to system deployers.

- *Art. 13(1) of the AI Act: Transparency and provision of information to deployers*
- *Art. 50 of the AI Act: Transparency obligations for system providers and deployers*

High-risk system deployers should verify that agent identification systems are properly implemented and maintained throughout their operational use. This includes ensuring that identifiers are used to clearly notify users when they are interacting with an agent, promoting transparency. Deployers should also regularly check that identifiers remain intact and accessible during the system's operation, ensure the accuracy and completeness of identifier information in the deployed environment, and maintain records of agent identifiers for all deployed AI systems under their responsibility.

- *Art. 26(1) of the AI Act: Technical and organisational measures*
- *Art. 50 of the AI Act: Transparency obligations for system providers and deployers*

4.2.2 Real-Time Monitoring

Real-time monitoring enables providers and tool developers to gain live insights into agentic activities by delivering automated alerts when abnormal or unauthorised actions are detected. Analogous to processes in modern Anti-Money Laundering (AML) systems, this could involve defining specific indicators around agent activity, which upon exceeding a predetermined threshold, would trigger an alert. In AML, indicators such as quantity and speed of transaction activity are tracked and flagged (FATF, 2020).

Similarly, agent-focused monitoring systems, built on agent infrastructure (A. Chan et al., 2025) and tailored to each agent's use case, could monitor several key variables that can act as indicators of risk. These variables might include (i) the number of agents involved in the operation, (ii) the scale of compute resources being used, (iii) the duration for which the agent operates autonomously without any human intervention, (iv) any economic transactions the agent initiates, (v) any potentially concerning interactions agents have with other agents, people, or businesses, and (vi) the usage of sensitive information (like logins or biometric data).

The risk thresholds for these variables are not fixed; they can be flexible and continuously updated. By employing existing tuning and optimisation methods, similar to those used in AML (Dettmer, 2024), the risk thresholds can be adjusted over time. This ensures the monitoring system remains responsive to changing conditions and can more accurately detect anomalies or potential risks as the system evolves. Once these thresholds are hit, developers are able to intervene, review, and potentially alert regulators, allowing for early detection and intervention when risks materialise. Additionally, monitoring could also incorporate information from multiple agents to capture multi-agent risks such as collusion. These approaches are still quite nascent, and still need to be validated, though labs such as Anthropic (2024a) have made a first advance on real-time (and asynchronous) monitoring. In addition, these monitoring capabilities should be further linked to reporting obligations (both internal and external).

GPAISR providers should develop real-time post-deployment monitoring capabilities for their models when used in agentic applications, whether through APIs or other means, that can detect abnormal or unauthorised agent activities. Model providers should develop flexible threshold mechanisms that system providers can tune to different operational contexts and risk profiles.

- *Measure II.4.14 of the Code of Practice: Post-market monitoring*
- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*
- *Measure II.6.1(2) of the Code of Practice: Safety mitigations: Input and output monitoring*

High-risk system providers should implement and configure the post-market monitoring capabilities provided by model providers, establishing appropriate thresholds based on their specific operational requirements and risk assessment. They should integrate these monitoring systems with their broader security infrastructure and establish clear response protocols when thresholds are exceeded. System providers should contribute to monitoring system improvement by providing feedback on alert accuracy and effectiveness within their specific deployment context.

- *Art. 72 and Recital 155 of the AI Act: Post-market monitoring*

High-risk system deployers should use the configured monitoring capabilities and follow the usage instructions provided by system providers to oversee the agent, notify system and model providers of any detected systemic risks, and suspend its use if necessary.

- *Art. 26(5) of the AI Act: Operation-monitoring*

4.2.3 Activity Logs

Activity logs are a key tool for capturing agents' decision-making processes and their interactions with the external world. These logs document agent inputs, outputs, scaffolding used (i.e., API calls made) and their timestamps, and possibly internal reasoning chains, akin to explainability tools (Wei et al, 2023). The amount of detail can be proportional to the operational environment and the risk levels identified during risk modelling. By preserving a detailed record of activities, providers or developers can perform post-incident analyses, tracing harmful outcomes back to their root causes, and improving AI agents to prevent future harms. These logs are also vital when distributing liability among actors along the value chain when incidents cause harm.

GPAISR providers should leverage agent identification infrastructure built for post-market monitoring to develop and maintain detailed logging systems that capture agent activity for retrospective inspection that document all model inputs, outputs, API calls, and reasoning chains with appropriate timestamps. Providers should ensure logging mechanisms maintain appropriate privacy protections while supporting effective post-incident analysis.

- *Measure II.4.14 of the Code of Practice: Post-market monitoring*
- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*

High-risk system providers should implement the logging functionality provided by model providers and extend it to capture activity details most relevant to specific deployment risks. They should establish appropriate data retention policies aligned with their operational requirements and regulatory obligations.

- *Art. 72 and Recital 155 of the AI Act: Post-market monitoring*
- *Art. 19 of the AI Act: Automatically generated logs*
- *Art. 12(3) of the AI Act: Record-keeping*

High-risk system deployers should retain the automatically generated logs for a period appropriate to the intended purpose of the agent, of at least six months, unless legal requirements on personal data protection apply.

- *Art. 26(6) of the AI Act: Log-keeping*

4.2.4 Acceptable Use Policies

Acceptable Use Policies (AUPs) are documents or agreements that clearly and explicitly outline permitted and prohibited uses of certain technologies (Doherty et al., 2010). AUPs establish explicit boundaries around how GPAI models may be integrated into agent systems, specifying permissible capabilities, external tool access, and operational constraints.

For AI agents, these policies are particularly important as they help mitigate systemic risks arising from capabilities and affordances such as autonomous operation and API access by defining key safety parameters such as maximum duration of unsupervised operation, required oversight measures, and conditions requiring human intervention. Well-designed AUPs create accountability across the value chain while providing clear guidance for downstream implementation. As discussed in Section 3, AUPs, along with appropriate additional measures, may be used by GPAI model providers to exclude high-risk uses from the intended purposes.

GPAISR providers must clearly articulate in their AUPs how their models may be used in agent systems. These policies should explicitly state whether the provider intends their model to be used in agent applications and establish broad yet clear boundaries between permitted and prohibited agent deployments. GPAI providers should also outline the technical and operational support they will provide to downstream providers implementing these controls.

- *Annex XII 1(b) of the EU AI Act: Technical documentation: Acceptable Use Policies*
- *Commitment I.1 of the Code of Practice: Model Documentation Form: Acceptable Use Policies*

High-risk system providers must thoroughly review and adhere to the AUPs established by model providers, ensuring their agent implementation complies with all specified constraints. They should also develop supplementary policies or voluntary model terms that address use-case specific considerations not covered in the general AUP provided by the model provider.

- *Art. 25 of the AI Act: Responsibilities along the AI value chain: Voluntary model terms*

High-risk system deployers should adhere to both the AUPs established by the model provider and the supplementary AUP policies on use-case specific considerations made by the system providers.

- *Art. 26(1) of the EU AI Act: Instructions for use*

4.2.5 Challenges and Future Directions

While transparency tools provide essential visibility into agent behavior, a key tension exists between visibility and privacy rights. As agents increasingly substitute for humans across various activities, information about their operations could effectively become surveillance data of personal activities (Goodwin, 2018). This creates a fundamental conflict: GPAI model providers are responding with privacy assurances, including APIs with no logging capabilities and guarantees against data retention (Hoder et al., 2024), which aligns with data protection frameworks like GDPR that regulate the collection of personal data (Finck and Pallas, 2020). This privacy-visibility trade-off creates blind spots that can prevent stakeholders from fully understanding or controlling agent decisions, particularly in high-risk environments. The resulting dilemma requires balancing transparency needs with privacy protection to avoid undue surveillance while maintaining adequate oversight.

Potential solutions include data trusts (Delacroix and Lawrence, 2019) and differentiated access controls that vary in granularity and quantity of information shared. For instance, regulators might receive detailed logs for high-risk activities, while other stakeholders may only be granted access to aggregated statistics (Bluemke et al., 2023). This challenge ties into broader questions of responsibility along the AI value chain, where different stakeholders, such as model or system providers or deployers, have varying rights to access information.

Under frameworks like the GDPR (European Parliament and Council, 2016), it may be difficult to justify granting model providers access to detailed data collected at the system deployer level, particularly if such data includes sensitive user interactions.

Technical challenges also persist, such as agents potentially circumventing monitoring by mimicking human behavior in their tool interactions. While CAPTCHA-like systems and identity verification protocols offer partial solutions (Egan and Heim, 2023), these measures must evolve alongside advancing agent capabilities. Future research should explore measures that enable human verification without compromising privacy, particularly for high-risk domains where strong identity assurance is crucial.

4.3 Technical Deployment Controls

4.3.1 Real-Time Action Refusals

Real-time action refusals build upon established content filtering approaches used in GPAI models. While these filters are typically implemented in the underlying models, their effectiveness requires special consideration of agent-specific risk dynamics.

Current best practices suggest implementing a risk severity framework that categorises different types of outputs and assigns appropriate filtering policies (Farley et al., 2024). Due to agents' ability to make multiple sequential decisions and interact with external systems, action refusals must account for both individual outputs and their cumulative effects. This requires implementing multi-level filtering that considers not only immediate content risks (such as harmful or unlawful content) but also potentially problematic patterns of behaviour that emerge over multiple interactions (Feng et al., 2017; Ji et al., 2024). Regular testing and updates are essential to maintain filter effectiveness while adapting to new patterns of agent behaviour, including systematic monitoring of user complaints and red-team testing to identify potential weaknesses (Shah, 2024).

GPAISR providers should enhance their existing real-time action refusal systems to address the unique challenges posed by agentic applications. Providers should develop multi-level filtering capabilities that can assess both individual outputs and their cumulative effects, addressing the unique risks posed by agents' ability to make multiple interconnected decisions.

- *Commitment II.6.1(2) of the Code of Practice: Safety mitigations: Output filtering*

- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*

High-risk system providers should implement and configure the general filtering capabilities provided by model providers, supplementing them with domain-specific filters that address risks unique to their particular deployment environment. Providers should establish monitoring protocols to evaluate filter effectiveness and contribute to ongoing improvement through systematic feedback on false positives and missed harmful content.

- *Art. 9(5) of the AI Act: Risk management measures*

High-risk system deployers should review and provide feedback on filter performance in operational contexts, ensuring filters adequately mitigate risks specific to their deployment environment. They should monitor for false positives that impact legitimate use and document instances where filtering fails to capture potentially harmful outputs, contributing to the continuous improvement of filtering systems.

- *Art. 26(1) & (5) of the EU AI Act: Technical and organisational measures, Operation-monitoring*

4.3.2 Emergency Shutdowns

Building upon the real-time monitoring infrastructure described before, automated shutdown mechanisms serve as a critical last-resort control measure for AI agents. When monitoring alerts indicate potentially dangerous behavior such as unusual API call patterns, excessive resource usage, or suspicious interactions with other agents or actors, these mechanisms can automatically pause or terminate the agent's operations (Hadfield-Menell et al., 2017). In addition to automated shutdowns, manual shutdown options should also be made available, allowing authorised personnel to halt operations independently of automated systems. These are crucial for detecting risks missed by automation or when monitoring fails. Shutdown controls should be accessible and separate from the agent's main system.

The shutdown process, drawing on established practices in industrial safety systems, must ensure swift yet controlled termination that prevents cascading failures in connected systems (Tajuddin, 2024). This includes immediately suspending external API access, safely terminating ongoing processes, and securely preserving system state for post-incident analysis. They should also include a limited functionality fallback mode to keep critical functions operating. As the AML systems mentioned before, shutdown thresholds can be dynamically adjusted based on operational experience and risk levels.

GPAISR providers should develop automated emergency shutdown capabilities and infrastructure that integrate with their real-time (post-deployment) monitoring systems, made available to other (downstream) actors. If such a shutdown is triggered, this must be reported immediately to the AI Office and other affected or relevant stakeholders.

- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*

High-risk system providers should include functionality for system deployers to be able to intervene or interrupt system operations. System providers should thus implement the emergency shutdown capabilities provided by model providers, integrating them with their broader system architecture and operational procedures. They should establish graduated response protocols based on alert severity, with clear criteria for triggering partial or complete shutdowns. They should develop comprehensive post-shutdown investigation procedures to identify root causes and implement appropriate remediation before system reactivation.

- *Art. 14(4)(d) of the AI Act: Human oversight*

High-risk system deployers should test operational protocols in response to emergency shutdown alerts, train staff on appropriate intervention measures, and maintain records of shutdown incidents for regulatory compliance and system improvement.

- *Art. 26(2) & (5) of the EU AI Act: Human oversight, Suspension of use*
- *Art. 4 of the AI Act: AI literacy*

4.3.3 Challenges and Future Directions

Regarding real-time action refusals, the multi-layered nature of AI agents introduces distinct filtering challenges. As agents currently rely on chain-of-thought processes (Wei et al., 2023) with multiple LLM calls, and use model output as input to external tools, like a web browser, filtering mechanisms must evolve beyond evaluating individual final outputs to consider how seemingly benign instructions might combine or interact to produce harmful outcomes (Feng et al., 2017; Shah, 2024). This requires sophisticated and insofar undeveloped approaches capable of anticipating how model outputs might manifest differently when interpreted and executed within the broader system context.

Maximising control effectiveness for emergency shutdowns can have an undesirable effect on operational continuity. Overly sensitive filters or shutdown mechanisms might disrupt legitimate agent operations and cause serious financial and logistical issues, while overly permissive ones could fail to prevent harmful outcomes.Technical integration challenges also persist, particularly in ensuring reliable operation of shutdown mechanisms across diverse deployment environments, which can be hard to foresee and implement at the model level.

Future research directions could also focus on investigating more nuanced shutdown procedures that can selectively restrict agent capabilities rather than requiring complete termination, and fail-safe mechanisms which ensure agent shutdowns cause minimal damage. This includes agent infrastructure for rollback mechanisms for incident response, both during normal operations, as well as specifically post-emergency shutdown

4.4 Human Oversight

Human-centric design should be used as a core design principle, tightly integrating controls into AI agents. This aligns conceptually with the human oversight provisions outlined in Art. 14 of the AI Act, which emphasises the need for meaningful human control in high-risk AI systems to ensure their safe operation and mitigate risks.

One major control design choice is human-in-the-loop (HITL), which is an established practice in the field of autonomous vehicles and is governed by the industry-standard SAE Levels of Driving Automation (SAE International, 2021), made in partnership with the International Organization for Standardization (ISO). SAE Levels of Driving Automation establish safety standards that vehicles should meet before being able to integrate varying degrees of automation. For instance, under normal situations and certain levels of driving automation, advanced driver assistance systems (ADAS) are permitted to take certain actions such as acceleration, braking, and steering, but drivers are expected to be on standby and take control if necessary.

While automation bias remains a concern in HITL systems (Goddard et al., 2012), they help maintain clear lines of accountability and can effectively prevent accidental harms by ensuring human judgment remains part of the decision-making process, facilitating compliance with possible future liability laws (Smakman et al., 2024).

4.4.1 Checkpoint System

The implementation of HITL controls involves setting specific checkpoints in the agent's workflow where human authorisation is required. These checkpoints can be triggered by various conditions, perhaps based on the transparency indicators of real-time monitoring previously outlined, including quantitative thresholds such as duration of unsupervised activity or number of API calls, as well as qualitative risk indicators like attempts to access high-risk APIs or requests for permissions outside the agent's predefined scope. These checkpoints serve as friction points in the agent's operation, automatically pausing further actions until human review is completed, mirroring established practices in banking and financial security where suspicious account activity triggers automatic suspensions pending review.

GPAISR providers should build foundational checkpoint infrastructure into their APIs, enabling automated pauses based on key risk indicators like rapid API calls, access attempts to sensitive resources, or suspicious outputs. This infrastructure should include default configurations, logging capabilities, and emergency shutdown options.

- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*

High-risk system providers should then customise this infrastructure for their specific use case by setting appropriate thresholds, defining clear human oversight protocols (including who can review or approve actions), and adding domain-specific checkpoints.

- *Art. 14 of the AI Act: Human oversight*

High-risk system deployers should ensure their staff have the requisite AI literacy, comprising elements such as technical knowledge, experience, and training to make informed and less biased decisions when reviewing actions at agentic checkpoints.

- *Art. 4 of the AI Act: AI literacy*

4.4.2 Permission Management

The design of agent permission systems can draw valuable lessons from mobile operating system security models. Modern smartphone platforms have evolved sophisticated permission frameworks that balance functionality with security, as exemplified by Android's shift to the ask-on-first-use (AOFU) model in 2015, which requires explicit user consent before applications can access sensitive resources (Lavranou et al., 2023). Applied to AI agents, this approach includes three key measures: (i) implementing explicit declaration of required permissions in the agent's configuration, (ii) dynamic permission requests at runtime for sensitive operations, and (iii) granular control over access to different APIs and system resources, all supported by clear documentation of permission implications and risks.

Leveraging agent identification infrastructure, providers can implement unique delegation credentials that explicitly define an agent's permitted actions and operational boundaries, specifying its identity, delegating user, and precise access limitations (South et al., 2025). This approach ensures precise control over agent capabilities while providing flexibility for legitimate operations. Additionally, dynamic permission control with real-time verification, such as cross-agent credential validation and periodic token rotation, can further mitigate security risks (South et al., 2025).

GPAISR providers should develop permission management systems for agent applications that enable granular control over agent capabilities and resource access. Providers should develop clear documentation regarding types of permissions and their relevant risks to support informed downstream deployment decisions.

- *Recital 101 and Annex XII(1)(d) of the AI Act: Transparency*
- *Measure II.6.1(7) of the Code of Practice: Agentic infrastructure*
- *Commitment I.1 of the Code of Practice: Documentation*

High-risk system providers should implement the permission management systems provided by model providers, configuring them to align with their specific operational requirements and risk profile.

- *Art. 14 of the AI Act: Human oversight*

High-risk system deployers should review and provide feedback on permission configurations based on operational experience.

- *Art. 26(5) of the EU AI Act: Operation-monitoring*
- *Art. 27(1)(e): Fundamental rights impact assessment: Human oversight implementation*

4.4.3 Challenges and Future Directions

A key challenge in implementing effective human oversight lies in identifying meaningful risk indicators and their relationship to potential harms. While established monitoring frameworks provide useful starting points, agents present unprecedented complexity in mapping activities to risks. The challenge extends beyond tracking individual permissions or actions to understanding how different agent activities might combine or cascade to create risk scenarios that are difficult to predict from monitoring single indicators alone. This requires developing a sophisticated understanding of which permissions and activities genuinely proliferate risk, rather than simply tracking conventional metrics that may not capture the unique ways agents can generate or amplify potential harms (Lavranou et al., 2023).

Regarding the checkpoint system, too many checkpoints can create significant bottlenecks in automated processes, potentially undermining the efficiency gains that agents promise. Conversely, too few might miss critical intervention points. This could include adaptive systems that adjust oversight requirements based on operational context and historical performance.

Human factors present additional challenges. Alert fatigue, a well-documented phenomenon in fields from healthcare to autonomous vehicle systems, can lead to perfunctory reviews and rubber-stamping of decisions (Cash, 2009; Goddard et al., 2012). This automation bias risk is particularly concerning in high-stakes domains where careful human judgment serves as a critical safety barrier.

5. Conclusion

AI agents have become integral to the strategic roadmaps of major model providers and downstream startups alike. Although current agents fail to reliably complete tasks across most domains, they have demonstrated strong capabilities in specialised areas like software engineering. If they live up to their promise, AI agents could have substantial economic impact, yet would also amplify AI risks, primarily through their capacity to engage in autonomous, long-term planning and directly influence real and virtual environments.

The EU AI Act was not explicitly designed with autonomous AI agents in mind. Nevertheless, its approach of regulating AI along the value chain, allocating responsibilities to model providers, system providers, and system deployers, appears promising for governing AI agents. Because most agents integrate a GPAISR, Ch. V obligations (Articles 53–55) apply, requiring providers to assess and mitigate the systemic risks arising from their models' use in AI agents. Specific AI agents represent an AI system under the EU AI Act. Additionally, we may also consider them a GPAI system if they use a GPAI model and can serve a variety of purposes. An agent is a high-risk system if it is used as a safety component or in a high-risk area per Annex III, resulting in the applicability of Ch. III to providers and deployers.

The AI Act governs agents through four primary pillars: risk assessment, transparency tools, technical deployment controls, and human oversight. We derive these complementary pillars by conducting an integrative review of the AI governance literature and mapping the results onto the EU AI Act. These measures must cascade through the entire value chain, with each actor playing distinct yet complementary roles suited to their positional expertise, resources and information. In general, model providers must build the fundamental infrastructure, system providers must adapt these tools to their specific context, and deployers must adhere to and apply these rules during operation.

However, several critical challenges persist across three main areas. First, further technical work is urgently needed to ensure the effectiveness of these mitigations: for instance, the agent evaluation ecosystem remains nascent, cross-sector shutdown mechanisms are technically complex, monitoring systems raise significant privacy implications, and human oversight remains vulnerable to automation bias. Second, potential gaps in the technical standards for Ch. III require revision, as current standards, such as human oversight requirements associated with Art. 14, are being developed for less autonomous AI systems and may fail to address the novel challenges posed by AI agents. Third, legal uncertainty regarding the applicability of the AI Act continues to present obstacles, particularly concerning the risk classification of GPAI systems and what it means to deliberately exclude high-risk uses. It also appears necessary to review and potentially amend high-risk uses under Annex III in light of AI agents.

Acknowledgments

We thank the following individuals for their helpful conversations, comments, and feedback on drafts of this policy briefing:

- Alan Chan (Centre for the Governance of AI)
- Alejandro Tlaie Boria (SaferAI)
- Afek Shamir (Pour Demain)
- Carson Ezell (Harvard University)
- Lewis Hammond (University of Oxford/Cooperative AI Foundation)
- Lisa Soder (Interface)
- Marta Bieńkiewicz (Cooperative AI Foundation)
- Merlin Stein (University of Oxford)
- Noam Kolt (Hebrew University)
- Saskia Welsch (SAP)
- Yohan Mathew (Centre for the Governance of AI)

A special thank you to all our incredible colleagues at The Future Society for supporting this project in a myriad of ways.

6. References

Allman, B. (2024). *6.3. Risk Identification and Evaluation*. Fanshawe College Pressbooks. https://ecampusontario.pressbooks.pub/projmgmteventplanning/chapter/6-3-risk-identification-and-evaluation/

Altman, S. (2025, January 6). *Reflections*. Sam Altman. https://blog.samaltman.com/reflections

Andriushchenko, M., Croce, F., & Flammarion, N. (2024a). Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks (No. arXiv:2404.02151). arXiv. https://doi.org/10.48550/arXiv.2404.02151

Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A., Kolter, Z., Fredrikson, M., Winsor, E., Wynne, J., Gal, Y., & Davies, X. (2024b). *AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents* (No. arXiv:2410.09024). arXiv. https://doi.org/10.48550/arXiv.2410.09024

Anthropic. (2024a, October 15). *Anthropic's Responsible Scaling Policy*. https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf

Anthropic. (2024b, October 22). *Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku*. https://www.anthropic.com/news/3-5-models-and-computer-use

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

AutoGPT. (2024, November 14). OpenAI's New AI Agent 'Operator' Could Change How You Work. *AutoGPT*. https://autogpt.net/openais-new-ai-agent-operator-could-change-how-you-work-heres-what-you-need-to-know/

Axios. (2025, January 23). *2025 is the year of AI agents, OpenAI CPO says*. https://www.axios.com/2025/01/23/davos-2025-ai-agents

Bluemke, E., Collins, T., Garfinkel, B., & Trask, A. (2023). *Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases* (No. arXiv:2303.08956). arXiv. https://doi.org/10.48550/arXiv.2303.08956

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A., & Schwaller, P. (2023, October 28). *Augmenting large language models with chemistry tools*. NeurIPS 2023 AI for Science Workshop. https://openreview.net/forum?id=wdGIL6lx3l

Cash, J. J. (2009). Alert fatigue. *American Journal of Health-System Pharmacy*, *66*(23), 2098–2101. https://doi.org/10.2146/ajhp090181

Casper, S., Bailey, L., Hunter, R., Ezell, C., Cabalé, E., Gerovitch, M., Slocum, S., Wei, K., Jurkovic, N., Khan, A., Christoffersen, P. J. K., Ozisik, A. P., Trivedi, R., Hadfield-Menell, D., & Kolt, N. (2025). *The AI Agent Index* (No. arXiv:2502.01635). arXiv. https://doi.org/10.48550/arXiv.2502.01635

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., … Hadfield-Menell, D. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. *The 2024 ACM*

Conference on Fairness, Accountability, and Transparency, 2254–2272. https://doi.org/10.1145/3630106.3659037

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 958–973. https://doi.org/10.1145/3630106.3658948

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., … Maharaj, T. (2023). Harms from Increasingly Agentic Algorithmic Systems. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666. https://doi.org/10.1145/3593013.3594033

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). *Infrastructure for AI Agents* (No. arXiv:2501.10114). arXiv. https://doi.org/10.48550/arXiv.2501.10114

Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., & Mądry, A. (2024). *MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering* (No. arXiv:2410.07095). arXiv. https://doi.org/10.48550/arXiv.2410.07095

China, C. R., & Goodwin, M. (2023, November 29). *What is API Monitoring?* IBM. https://www.ibm.com/think/topics/api-monitoring

Cihon, P. (2024). Chilling autonomy: Policy enforcement for human oversight of AI agents. *ICML*. https://blog.genlaw.org/pdfs/genlaw_icml2024/79.pdf

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding accountability in algorithmic value chains. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1186–1197. https://doi.org/10.1145/3593013.3594073

Delacroix, S., & Lawrence, N. D. (2019). Bottom-up data Trusts: Disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law*, *9*(4), 236–252. https://doi.org/10.1093/idpl/ipz014

Dettmer, R. (2024, July 10). *System Tuning and Optimization: The Key to Financial Compliance Longevity*. https://www.amlrightsource.com/news/system-tuning-and-optimization-the-key-to-financial-compliance-longevity

Doherty, N. F., Anastasakis, L., & Fulford, H. (2011). Reinforcing the security of corporate information resources: A critical review of the role of the acceptable use policy. *International Journal of Information Management*, *31*(3), 201–209. https://doi.org/10.1016/j.ijinfomgt.2010.06.001

Drexler, K. E. (2019). *Reframing Superintelligence*.

Drope, L. (2024, November 26). *How Do AI Agents Work? A Beginner's Guide*. https://www.lindy.ai/blog/how-do-ai-agents-work

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L., & Gao, J. (2024). *Agent AI: Surveying the Horizons of Multimodal Interaction* (No. arXiv:2401.03568). arXiv. https://doi.org/10.48550/arXiv.2401.03568

Egan, J., & Heim, L. (2023). *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers* (No. arXiv:2310.13625). arXiv. https://doi.org/10.48550/arXiv.2310.13625

European Commission. (2025, February 6). *The Commission publishes guidelines on AI system definition to facilitate the first AI Act's rules application | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application

European Parliament & European Council. (2016). *General Data Protection Regulation*. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Fang, R., Bindu, R., Gupta, A., Zhan, Q., & Kang, D. (2024). *LLM Agents can Autonomously Hack Websites* (No. arXiv:2402.06664). arXiv. https://doi.org/10.48550/arXiv.2402.06664

Farley, P., Gilley, S., Urban, E., Jenks, A., Bullwinkle, M., Lakkoju, M., Bradish, D., & Greene, M. (2024, August 28). *Content Filtering*. Azure OpenAI Service Documentation. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter

Feinstein, D. (2019, July 16). *S.2125 - 116th Congress (2019-2020): Bot Disclosure and Accountability Act of 2019* (2019-07-16). Congress. https://www.congress.gov/bill/116th-congress/senate-bill/2125

Feng, C., Li, T., & Chana, D. (2017). Multi-level Anomaly Detection in Industrial Control Systems via Package Signatures and LSTM Networks. *2017 47th Annual IEEE/IFIP*

International Conference on Dependable Systems and Networks (DSN), 261–272. https://doi.org/10.1109/DSN.2017.34

Financial Action Task Force (FATF). (2020). *Virtual Assets Red Flag Indicators of Money Laundering and Terrorist Financing Financial and Non-Financial Sectors*. https://www.fatf-gafi.org/content/dam/fatf-gafi/brochures/Handout-Red-Flags-VA-Financial-Non-Financial.pdf

Finck, M., & Pallas, F. (2020). They who must not be identified—Distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law*, *10*(1), 11–36. https://doi.org/10.1093/idpl/ipz026

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., … Manyika, J. (2024). *The Ethics of Advanced AI Assistants* (No. arXiv:2404.16244). arXiv. http://arxiv.org/abs/2404.16244

Gao, K., Vytelingum, P., Weston, S., Luk, W., & Guo, C. (2024). High-frequency financial market simulation and flash crash scenarios analysis: An agent-based modelling approach. *Journal of Artificial Societies and Social Simulation*, *27*(2), 8. https://doi.org/10.18564/jasss.5403

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goodwin, C. (2018, April 10). *Cooperation or Resistance?: The Role of Tech Companies in Government*. Harvard Law Review.

https://harvardlawreview.org/print/vol-131/cooperation-or-resistance-the-role-of-tech-companies-in-government-surveillance/

Google. (2024, December 11). *Introducing Gemini 2.0: Our new AI model for the agentic era*. Google. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). *The Off-Switch Game* (No. arXiv:1611.08219). arXiv. https://doi.org/10.48550/arXiv.1611.08219

He, Y., Wang, E., Rong, Y., Cheng, Z., & Chen, H. (2024, June 20). *Security of AI Agents*. https://arxiv.org/html/2406.08689v2#S3

Heylighen, F., & Joslyn, C. (2003). Cybernetics and Second-Order Cybernetics. In R. A. Meyers (Ed.), Encyclopedia of Physical Science and Technology (Third Edition) (pp. 155–169). Academic Press. https://doi.org/10.1016/B0-12-227410-5/00161-7

Hoder, C., Farley, P., Urban, E., Mehrotra, N., Bullwinkle, M., & Hill, A. (2024, December 18). *Data, privacy, and security for Azure OpenAI Service—Azure AI services*. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy

Hugging Face. (n.d.). *Model Cards*. Retrieved 6 January 2025, from https://huggingface.co/docs/hub/en/model-cards

International Network of AI Safety Institutes. (2024). *Improving International Testing of Foundation Models- A Pilot Testing Exercise from the International Network of AI Safety Institutes*. https://www.nist.gov/system/files/documents/2024/11/21/Improving%20International%20Testing%20of%20Foundation%20Models-%20%20%20A%20Pilot%20

Testing%20Exercise%20from%20the%20International%20Network%20of%20AI%20Safety%20Institutes.pdf

Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Li, B., & Yang, Y. (2024). *PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference* (No. arXiv:2406.15513). arXiv. https://doi.org/10.48550/arXiv.2406.15513

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. R. (2023, October 13). *SWE-bench: Can Language Models Resolve Real-world Github Issues?* The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=VTF8yNQM66

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). *AI Agents That Matter* (No. arXiv:2407.01502). arXiv. http://arxiv.org/abs/2407.01502

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., Ho, A., Barnes, E., & Christiano, P. (n.d.). *Evaluating Language-Model Agents on Realistic Autonomous Tasks*.

Kolt, N. (2025). Governing AI Agents. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4772956

Küspert, S., Moës, N., & Dunlop, C. (2023, February 10). *The value chain of general-purpose AI*. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/

Kraprayoon, J., Williams, Z., & Fayyaz, R. (2025). AI Agent Governance: A Field Guide. Institute for AI Policy and Strategy (IAPS). https://www.iaps.ai/research/ai-agent-governance

Lavranou, R., Karagiannis, S., Tsohou, A., & Magkos, E. (2023). Unraveling the Complexity of Mobile Application Permissions: Strategies to Enhance Users' Privacy Education. *European Journal of Engineering and Technology Research*, 87–95. https://doi.org/10.24018/ejeng.2023.1.CIE.3141

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., … Tang, J. (2023). *AgentBench: Evaluating LLMs as Agents* (No. arXiv:2308.03688). arXiv. https://doi.org/10.48550/arXiv.2308.03688

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). *Frontier Models are Capable of In-context Scheming* (No. arXiv:2412.04984). arXiv. https://doi.org/10.48550/arXiv.2412.04984

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). *GAIA: A benchmark for General AI Assistants* (No. arXiv:2311.12983). arXiv. https://doi.org/10.48550/arXiv.2311.12983

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Moës, N., & Ryan, F. (2023). Heavy is the Head that Wears the Crown: A risk-based tiered approach to governing General-Purpose AI. The Future Society. https://thefuturesociety.org/heavy-is-the-head-that-wears-the-crown/

Montgomery, B. (2024, October 23). Mother says AI chatbot led her son to kill himself in lawsuit against its maker. *The Guardian*.

https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., & Raji, I. D. (2024). *Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling* (No. arXiv:2402.17861). arXiv. http://arxiv.org/abs/2402.17861

OpenAI. (2025a, January 23). *Introducing Operator*. https://openai.com/index/introducing-operator/

OpenAI. (2025b, February 2). *Introducing deep research*. https://openai.com/index/introducing-deep-research/

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., … Trager, R. (2024). *Open Problems in Technical AI Governance* (No. arXiv:2407.14981). arXiv. http://arxiv.org/abs/2407.14981

Russell, M., & Stokes, S. (2025, January 9). Silicon Valley is licking its chops at the promise of AI 'agents.' These are the startups to watch. Business Insider. https://www.businessinsider.com/startups-ai-agents-raising-venture-funding-2025-1

SAE International. (2021, April 30). *J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. https://www.sae.org/standards/content/j3016_202104/

Salesforce. (2024, September 12). *Salesforce Unveils Agentforce–What AI Was Meant to Be—Salesforce*.

https://www.salesforce.com/news/press-releases/2024/09/12/agentforce-announcement/

SAP. (2024, December 19). What are AI agents? SAP. https://www.sap.com/hk/resources/what-are-ai-agents

Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). *Large Language Models can Strategically Deceive their Users when Put Under Pressure* (No. arXiv:2311.07590). arXiv. https://doi.org/10.48550/arXiv.2311.07590

Schwartmann, R., & Zenner, K. (2025). GPAI Applications Under Scrutiny: The Regulation of the AI Regulation Along the Value Chain. Journal for European Data and Information Law, 1, 3–9. https://www.nomos.de/zeitschriften/eudir/

Shah, D. (2024, May 15). *AI Under Siege: Red-Teaming Large Language Models*. Lakera. https://www.lakera.ai/blog/ai-red-teaming

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for Governing Agentic AI Systems*.

Shpigelman, F. (2024, July 10). *How a kitten easily bypassed Azure AI content filtering*. Apex. https://www.apexhq.ai/blog/blog/from-kittens-to-alignment-how-a-kitten-easily-bypassed-azure-ai-content-filtering/

Smakman, J., Soder, L., Dunlop, C., Pan, W., Swaroop, S., & Kolt, N. (2024). An Autonomy-Based Classification: AI Agents, Liability and learnings from the UK Automated Vehicles Act. *2nd Workshop on Regulatable ML (NeurIPS 2024)*. https://openreview.net/pdf?id=EH6SmoChx9

South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Greenwood, D., Chan, A., & Pentland, A. (2025). *Authenticated Delegation and Authorized AI Agents* (No. arXiv:2501.09674). arXiv. https://doi.org/10.48550/arXiv.2501.09674

Tajuddin, A. (2024, August 16). Emergency Shutdown Systems and Procedures—Safety Notes. *Safetynotes.Net*. https://www.safetynotes.net/emergency-shutdown-systems-and-procedures/

Toner, H., Bansemer, J., Crichton, K., Burtell, M., & Woodside, T. (2024). *Through the Chat Window and Into the Real World: Preparing for AI Agents*. Center for Security and Emerging Technology (CSET). https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/

UK AISI. (2024, October 24). *Early lessons from evaluating frontier AI systems*. https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. http://arxiv.org/abs/2201.11903

Weidinger, L., Barnhart, J., Brennan, J., Butterfield, C., Young, S., Hawkins, W., Hendricks, L. A., Comanescu, R., Chang, O., Rodriguez, M., Beroshi, J., Bloxwich, D., Proleev, L., Chen, J., Farquhar, S., Ho, L., Gabriel, I., Dafoe, A., & Isaac, W. (2024). *Holistic Safety and Responsibility Evaluations of Advanced AI Models* (No. arXiv:2404.14068). arXiv. http://arxiv.org/abs/2404.14068

Xu, R., Wang, Z., Fan, R.-Z., & Liu, P. (2024). *Benchmarking Benchmark Leakage in Large Language Models* (No. arXiv:2404.18824). arXiv. https://doi.org/10.48550/arXiv.2404.18824

ZDNET. (2025, January 30). *AI agents will match 'good mid-level' engineers this year, says Mark Zuckerberg*. ZDNET. https://www.zdnet.com/article/ai-agents-will-match-good-mid-level-engineers-this-year-says-mark-zuckerberg/

7. Appendix

7.1 Analysing the GAIA Benchmark

GAIA is one of several benchmarks to evaluate the performance of agents (i.e. General AI Assistants), composed of 466 real-world questions which require real-world questions that require a set of fundamental abilities such as reasoning, multi-modality handling, web browsing, and general tool-use proficiency (Mialon et al., 2024). Importantly, the GAIA leaderboard is public and invites developers of AI agents to test their systems and share the result. Currently, the leaderboard contains information on the performance of 43 agents. We excluded 6 agents since they represented different versions of the same agent which were released within a 1 month timeframe, solely keeping the latest version, as such redundancies undermine the assessment of trends in agent performance over time.

7.2 Applicability of the EU AI Act: Agents as Modified GPAI Models

Building an agent may constitute a modification of the GPAI model necessitating limited applicability of Ch. V obligations to the downstream provider. Recital 109 states that (underlining added): [...] In the case of a modification or fine-tuning of a model, the obligations for providers of general-purpose AI models should be limited to that modification or fine-tuning, for example by complementing the already existing technical documentation with information on the modifications, including new training data sources, as a means to comply with the value chain obligations provided in this Regulation." Analogous to Art. 3(23) a transfer of obligations appears appropriate when the modification is "substantial". Whereas the Act itself does not present any guidance in this regard, the recently issued consultation by the AI Office articulates an understanding of GPAI model modification that is limited to fine-tuning, suggesting that it requires the downstream provider to spend 3 10^23 FLOP on fine-tuning (European Commission, 2025). By contrast, a more expansive interpretation could assume that common agent scaffoldings, such as retrieval-augmented generation (RAG), or frameworks to organise API calls, may constitute substantial modifications to the model. When accepting such an interpretation, this would warrant the applicability of obligations under Ch. V to those specific modifications, in scenarios where the agent is developed by downstream providers.

THE
FUTURE
SOCIETY

Contact Us

GENERAL info@thefuturesociety.org
PRESS press@thefuturesociety.org

THE FUTURE SOCIETY

www.thefuturesociety.org