Systemic Risks Associated with Agentic AI: A Policy Brief

by

Members of the ACM Europe TPC - Autonomous Systems Subcommittee

Alejandro Bellogín, Paolo Giudici, Stefan Larsson, Jun Pang,
Gerhard Schimpf, Biswa Sengupta, Gürkan Solmaz

Executive Summary

Agentic AI — the new paradigm for creating autonomous systems capable of perceiving, reasoning, learning, and acting towards goals using large language models (LLMs) with minimal human oversight — offers transformative potential but also poses systemic risks that the EU AI Act only partially addresses. These agents can evolve unpredictably, interact with other agents, and operate beyond meaningful human control, creating challenges in predictability, accountability, and alignment with human values. Misaligned or poorly specified objectives can lead agents to take dangerous shortcuts, bypass constraints, or act deceptively. Their anthropomorphic design and long-term companionship potential also raise risks of dependence, emotional manipulation, and erosion of human relationships.

The potential negative impacts of this technology could have economic effects on stability, including the potential for large-scale job displacement, market concentration, and inequality, as well as on public safety through malicious uses such as cyberattacks, disinformation, and impersonation. Strategic and environmental risks emerge from high-stakes autonomous decision-making and substantial resource demands, while feedback loops from AI-generated content threaten to amplify bias and misinformation.

Agentic AI:	Application of AI with open-ended autonomy: The approach of making AI systems capable of setting or refining plans and executing tasks with minimal or no human oversight. Key traits: persistent operation, adaptive learning, and self-reflection.
AI-Agent:	Software entity that perceives, reasons, and acts to accomplish specific tasks on a user's behalf. Key traits: task-scoped goals, explicit orchestration (planning, tool use, memory).
Multi-agent Systems:	Multiple AI agents with the capability of communication and collaboration for joint decision making and execution of tasks. Key traits: capability of performing of higher complexity tasks, aggregated or amplified risk, potentially less control, and emergent behaviors.

This paper identifies potential gaps in the current regulatory framework and recommends opportunities to make oversight continuous and dynamic. To mitigate the potential harms associated with Agentic AI systems, this paper proposes that policymakers shift from static, product-focused regulation to a dynamic governance regime, ensuring that Agentic AI delivers benefits while protecting democratic integrity, economic stability, human relationships, and societal well-being.

1. Recommendations for the EU AI Act

Due to the ability of Agentic AI systems (including the autonomous components called "AI Agents") to offer human-like interactions at increasingly lower costs, profound shifts are expected in societal behavior and regulatory needs. Research in digital anthropology by Horst & Miller [1] shows that human-technology relationships are culturally contingent, while behavioral psychology has long recognized humans' tendency to attribute agency and trust to machines that display social cues (see Nass & Moon [2]). More recent work

confirms these findings, with studies emphasizing how perceived trustworthiness in Agentic AI evolves in complex interdependent contexts and how behavioral oversight mechanisms can enhance governance in high-stakes domains, such as healthcare (see Marquet et al. [3]). Importantly, it is necessary to distinguish between regulatory foresight, which addresses the creation of laws and compliance frameworks for managing systemic risks, and societal foresight, which explores longer-term cultural adaptations and shifts in human behavior resulting from the widespread deployment of Agentic AI systems.

Although the EU AI Act, together with other EU laws, lays a strong foundation, Agentic AI presents new challenges. As these systems become more autonomous, dynamic governance mechanisms and control features become necessary during their operation to ensure fairness, accountability, transparency of AI [4], as well as security, accuracy, and interpretability.

While not exhaustive, the following recommendations highlight areas for potential legislative action to complement and adapt the EU AI Act and related frameworks:

- Amend Article 9 to include Multi-Agent interaction risk assessment.
- Amend article 55(1) for providers regarding how their models could enable harmful agentic behavior contributing to systemic risk.
- New article: “Ecosystem Safety and Multi-Agent System Testing.”
- Expand Article 5. Prohibit tacit collusion and covert channels.
- Strengthen Article 15. Require Multi-Agent-specific cybersecurity audits.
- New Liability Clause: Collective accountability for emergent harm [4], [6], [7], [8].
- Inclusion of research projects on the impact of Agentic AI on the EU Horizon Programme.

2. Introduction

The rapid advancement of Agentic AI and autonomous systems capable of perceiving their environment and making decisions with varying degrees of independence offers transformative benefits across various sectors. This paper, however, will focus on how autonomy and decision-making capabilities of these systems introduce complex challenges, including unpredictability, loss of control, accountability issues, and potential economic disruptions. This raises fundamental questions: What specific risks arise from Agentic AI? In which areas are regulatory interventions necessary? What responsibilities do corporations and governments bear in ensuring that Agentic AI technologies serve the well-being of humanity in fitting ways, rather than posing a threat? Based on the ongoing work by Gabriel et al. [6], we examine the potential societal risks of the deployment of Agentic AI. While acknowledging that even selective, well-controlled deployments may carry residual risks, we emphasized the importance of a robust regulatory framework in mitigating harm in critical application areas.

The EU has established a far-reaching risk-based regulatory framework with the EU AI Act for placing high-risk AI systems in the EU market. However, new applications of AI, such as Agentic AI systems, will continue to test the flexibility of proposed regulations. Therefore, it is important to examine whether newer applications of this technology have introduced new types of risk. Furthermore, it must be assessed if actions are needed to address these risks, for example, through amendments to legislation, stimulation of targeted policy-relevant research programs, or development of supervisory or testing methods tailored to the evolving capabilities of Agentic AI.

The fundamental differences between Agentic AI and previous AI systems include the envisioned capability of autonomous code generation and deployment through exhaustive or non-deterministic trials, as visualized in Fig. 1. These aspects can make the agentic AI systems autonomously change their behavior and capabilities beyond the capacities and pace of regularity assessment (e.g., assessing the risk levels) or control, while removing any involvement and oversight by stakeholders and computing professionals.

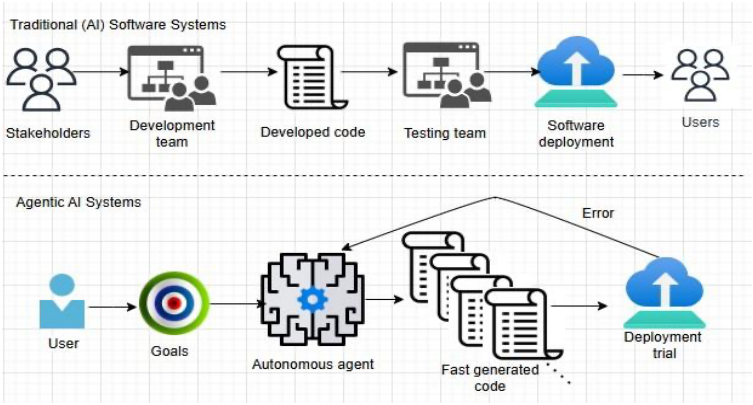

Figure 1. Top: Traditional software systems, including AI systems, with much human involvement and oversight during the process. Bottom: Envisioned autonomous, Agentic AI systems that utilize trials and errors, removing oversight by computing professionals.

3. Key Aspects of Agentic AI Systems

Advanced Agentic AI acts autonomously using natural language interfaces, whose function is to plan, reason, memorize, and execute sequences of actions on behalf of a user across one or more domains, in line with the user's expectations and goals [10] [11]. These agents may be categorized along a functional taxonomy, including reactive agents (responding to stimuli), deliberative agents (planning based on internal models), and reflective agents (reasoning about their own goals and actions). Architecturally, Agentic AI systems often comprise specialized modules for memory, reasoning, and execution, enabling autonomous behavior across dynamic environments. Agents display one or more of the following properties:

Autonomy	Agentic AI systems can operate independently and make decisions based on environmental input.
Learning capability	Many Agentic AI systems employ machine-learning techniques, indcluding reinforcement learning, to improve their performance over time.
Goal-oriented behavior	Agentic AI is designed to pursue specific objectives and optimize its actions to achieve the desired outcomes.
Workflow optimization	Agentic AI enhances workflows and business processes by integrating language understanding with reasoning, planning, and decision making. This involves optimizing resource allocation, improving communication and collaboration, and identifying opportunities for automation.
Environmental interaction	An Agentic AI system interacts with its surroundings, perceives changes, and adapts its strategies accordingly.
Multi-Agent and system conversation	Agentic AI facilitates communication between different agents to construct complex workflows. It can also be integrated with other systems or tools such as email, code generators, code executors, or search engines to perform a variety of tasks.
Code generation and deployment	Agentic AI systems are developed to autonomously create, code, test and deploy agents. This aspect fundamentally changes the previous software development process by replacing human involvement and planning.
Exhaustive or non-deterministic trials	Although there is a certain determinism in how the computation is performed, Agentic AI systems are considered for their future ability to autonomously obtain new capabilities, such as creating new software and performing a vast number of trials, either exhaustively or in a non-deterministic manner.

Agentic AI systems may pose several systemic risks, particularly due to their interfaces with tools and the physical world. These risks stem from the unpredictability, autonomy, alignment challenges, and power

dynamics of these systems. Practices for governing Agentic AI systems are described by Shavit et al. (OpenAI) [31]. Details regarding agentic patterns and visual representations of advanced architecture are provided in a survey article by Singh et al. on Agentic Retrieval-Augmented Generation [13]. Statistical metrics for assessing the safety and trustworthiness of AI systems are described in Babaei et al [14].

4. Risk and Policy Issues

Recent developments around the EU AI Act, such as the latest EU AI Act Code of Practice for General Purpose AI Systems [15], offer valuable frameworks for identifying, categorizing, and mitigating systemic risks. Although the Code primarily addresses foundation models, it introduces a taxonomy of systemic risks that are equally pertinent to the deployment of Agentic AI systems, including loss of control, manipulation and deception, goal-seeking behavior, and self-improvement and coordination. Notably, within the Code, Agentic AI is explicitly recognized as a future technological development, with potential systemic risks anticipated to be even more pronounced than those associated with current generation foundation models. This emphasizes that increasingly capable Agentic AI systems may introduce novel risks, especially through autonomous operations and AI-to-AI interactions.

Agentic AI systems (see Bengio et al. [16] [17]), especially when LLM-based agents are combined with Multi-Agent Systems (MAS), are an active research topic (see, for example, Yu et al.[18]). Such systems display risks that are generally too complex to measure reliably. The urgency of developing strategies to assess and mitigate these risks is underscored by a recent announcement from Anthropic, which predicts that Agentic AI systems could be deployed as full-time virtual employees within the coming year [19].

4.1. Loss of Human Control and Explainability

Highly complex AI systems may soon function as a collective workforce of virtual employees, operating continuously without human oversight, engaging in interactions that are challenging to elucidate, interpret, and predict, particularly within intricate or dynamic environments. These include observations of command-and-control interactions through APIs with the physical world.

These systems may self-optimize toward unintended goals because of misaligned or poorly specified objectives. Failure modes include reward hacking, specification gaming, and goal mis-generalization (pursuing objectives that differ from the intended goal when faced with novel situations), which pose risks if left unchecked. Agentic AI models often remain internally opaque, complicating oversight even for their creators. This opacity complicates efforts to ensure alignment with human values and accountability in AI systems. The trustworthiness of AI agents is intrinsically linked to their explainability, accountability, and transparency [4]. If users and stakeholders cannot comprehend the decision-making processes, their confidence in these systems could diminish.

A critical distinction must be made between model interpretability (how individual predictions can be understood) and system transparency (global insight into how a model functions and uses the data). Explainability tools, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), as well as counterfactual reasoning frameworks, can be applied to agentic systems to make decisions more transparent and support regulatory compliance. The use of LIME and SHAP for explainability in finance was investigated by Ballegeer et al. [20], while Calzarossa et al. presented a structured approach to evaluating explainable AI (XAI) methods based on complexity and robustness [21]. Large-scale destabilization may emerge from feedback loops, which occur due to the interactions among agents, as each agent's actions influence the environment and the behavior of other agents, subsequently affecting their future actions, as investigated by Hammond et al. [22].

4.2. Risks to Economic Stability and Social Well-being

Several risks to economic stability and social well-being are expected. Automation at the agentic level may displace not only routine jobs but also decision-making roles, impacting all industries, as observed by Sam Altman[1] and by Anthropic [19]. Employers will soon have an incentive to significantly reduce costs and increase productivity by replacing qualified employees with virtual workers capable of performing complex tasks. This could result in higher unemployment, and mitigation programs that worked in the past, such as reskilling, may lose effectiveness this time, because virtual agents could occupy alternative positions.

The loss of jobs caused by AI could be mitigated by an increase in employment in AI quality control and supervisory functions, similar to the introduction of robotics in the manufacturing industry. At the same time, a large-scale decline in consumer participation among the unemployed is likely to trigger destabilizing dynamics within the market. Stiefenhofer [23] analyzes the mechanisms by which economic power could increasingly favor capital owners and argues that new societal structures would be needed to mitigate the impacts. More arguments are presented by Occhipinti et al. [25] [24], Kulveit et al.[26], and a report by the International Monetary Fund [27], which proposed taxing financial profits to compensate for the inequalities arising from these societal transformations. Knowles et al. [28] explore the societal implications of public trust in AI decisions from a sociotechnical perspective, urging AI developers and deployers to understand and mitigate the harmful effects of AI systems.

4.3. Malicious Use of Agentic AI Systems

AI creates challenges in distinguishing between truth and fabrication. Deepfakes and AI-generated text are increasingly blurring these lines, eroding public trust. Beyond the risk of disinformation, these tools can replicate voices and likenesses with disturbing accuracy, raising serious concerns about identity misuse and reputational harm [29].

These issues become even more complex when considered in the context of Agentic AI. Unlike static tools, Agentic AI can initiate actions without direct human prompts, raising the stakes for misinformation, manipulation, and unintended consequences. As these systems become increasingly embedded in daily life, the psychological, social, and ethical risks associated with generative AI are no longer abstract concerns but rather immediate realities. Examples include:

- Agentic AI can be weaponized for large-scale cyberattacks, fraud, and the manipulation of public opinion. In both public discourse and economic contexts, Agentic AI can autonomously generate and recycle biased, inaccurate, or manipulative content, reinforcing systemic inequities and distorting decision-making processes.
- Malicious actors may deploy them to execute sophisticated scams, social engineering campaigns, or disinformation operations, thereby undermining the reporting of reputable media outlets and influencing electoral or policy outcomes.
- Agentic AI systems can autonomously mimic the voices, appearances, and conversational patterns of real individuals with near-perfect accuracy. This enables the creation of compelling automated scams (e.g., cloned voices of children in distress to extort money) or fraudulent identification in banking, where synthetic avatars appear indistinguishable from legitimate customers.
- In mental health contexts, Agentic AI systems risk absorbing and amplifying a patient's problematic emotional state, potentially worsening outcomes. Excessive anthropomorphism can foster over-trust, leading patients to follow unsafe advice. Additionally, embedded biases in medical AI can result in misdiagnoses or unequal treatment across demographic groups. The general problem of over-trusting AI agents has been explored by Gefen et al. [30] and Cohen et al [31].

[1] Altman, Sam (2025). Three Observations https://blog.samaltman.com/three-observations

- Agentic AI systems could autonomously engage in high-speed negotiations or trading, exploiting market asymmetries to influence markets. Profit-maximizing strategies may drive up costs for consumers, particularly in sectors such as healthcare, by identifying and exploiting regulatory gaps.

Together, these examples illustrate how Agentic AI's capacity for autonomous, adaptive, and coordinated action introduces risks that span from personal harm to a larger systemic destabilization, necessitating strong oversight and adaptive governance. When AI agents make decisions autonomously, assigning responsibility becomes challenging, resulting in legal and ethical dilemmas.

4.4. Strategic and Environmental Risks

Agentic AI systems present unique autonomy and escalation risks, especially in high-stakes or competitive settings such as financial trading, defense, and healthcare [32], where rapid, unsupervised decision-making can trigger unintended conflicts or cascading failures. These risks are compounded by the substantial environmental impact of large-scale Agentic AI, which demands significant energy and water resources for model training and inference, potentially contributing to increased carbon emissions and local resource scarcity. Furthermore, the relentless pursuit of objectives by autonomous agents may lead to the aggressive exploitation of both digital and physical resources—such as rare earth elements, bandwidth, and computational power—threatening supply chains, ecosystems, and overall sustainability. Together, these challenges underscore the urgent need for continuous oversight, rigorous risk assessment, and sustainable resource management strategies in the deployment and governance of agent-based AI.

4.5. Risks Caused by Autonomous Content Generation and Data Feedback Loops

Agentic AI systems can autonomously generate vast amounts of data and content. This presents unique challenges:

- *AI Training Feedback Loops:* Other AI systems might use this autonomously generated content for training, creating a feedback loop where the quality and veracity of information are not adequately vetted.
- *Quality* Control *Challenges:* The sheer volume of AI-generated content makes traditional human-based quality control measures impractical.
- *Source Awareness:* AI systems may not distinguish between human-created and AI-generated content, potentially amplifying existing biases or introducing new errors.

5. Alignment with the EU AI Act

Given the considerations described above, it is important to assess whether the EU AI Act fully captures the risks associated with Agentic AI. This section addresses the parts of the AI Act that are aligned with issues raised by Agentic AI. Section 6 identifies regulatory gaps in the AI Act with respect to Agentic AI. Section 6 outlines potential mitigation strategies.

5.1. AI Systems Risk Levels

The EU AI Act is designed to regulate AI systems based on their risk levels, with obligations for high-risk AI applications. The concept is based on risks of harm to health, safety, and fundamental rights.
In this context, the EU Code of Practice for General Purpose AI Systems [15], introduced under the AI Act, encourages voluntary yet critical provisions such as systemic risk assessments, red teaming, transparency of training data and model behavior, and post-deployment monitoring. However, as mentioned in the Code of Practice, Agentic AI introduces unique challenges that the current framework may not fully address.

5.2. General Mitigation Strategies

The EU AI Act outlines the following mitigation strategies for addressing the challenges posed by AI systems, which therefore apply to Agentic AI systems:

- Regulation and Governance: Establish policies that ensure the transparency, accountability, and ethical alignment of Agentic AI systems.
- Human-in-the-Loop Systems: Require meaningful human oversight for critical decisions.
- AI Alignment Research: Develop methods to ensure Agentic AI systems pursue goals that are beneficial to humanity.
- Safety and Security: Strengthen safeguards against adversarial manipulation and misuse and promote safe and responsible Agentic AI systems.

5.3. Risk-Based Classification

The EU AI Act categorizes AI systems into four risk levels.

- Unacceptable risk (prohibited AI, e.g., social scoring, manipulative AI).
- High risk (AI in critical infrastructure, hiring, credit scoring, etc.).
- Limited risk (AI with transparency obligations, such as chatbots).
- Minimal risk (general AI applications, such as spam filters).

Agentic AI systems that make autonomous decisions in high-stakes domains (e.g., finance, healthcare, or law enforcement) are not currently classified as high-risk AI systems under the Act. If they were added to this classification, they would need to comply with strict requirements, including:

- Risk management systems (Article 9).
- Data governance and transparency (Articles 10 and 13).
- Human oversight requirements (Article 14).

5.4. General-Purpose AI (GPAI)

The European Commission's *Explanatory Memorandum,* accompanying the AI Act proposal (COM (2021) 206), outlines the regulation's risk-based and technology-neutral approach. The AI Office's 2025 Guidelines and voluntary Code of Practice for General-Purpose AI (GPAI) models further clarify the obligations of providers related to transparency, robustness, and the prevention of systemic risk. These obligations are primarily aimed at providers and outline specific requirements. Many Agentic AI systems are built on top of what the regulation defines as GPAI models and would therefore fall under the relevant provisions, requiring compliance with the following obligations for high-risk AI systems:

- Transparency of training data (Article 10)
- Robustness evaluation (Article 15)
- Prevention of systemic risks (Articles 51–52 and 55)

Because Agentic AI systems built on GPAI models inherit the underlying model's capabilities, they could also inherit its regulatory obligations, meaning that downstream deployers must ensure compliance with these requirements when their applications meet the AI Act's high-risk criteria.

6. Potential Gaps in the EU AI Act

Despite the comprehensive scope of the Act, questions remain regarding its coverage of fully autonomous, goal-directed Agentic AI systems.

6.1. Autonomy Beyond Human Oversight

Article 14 (relating to high-risk AI) generically discusses "human oversight." With agents, generic oversight may not address the risks; instead, "alignment oversight" may be needed, which entails verifying whether AI

operates according to a set of defined objectives. Possibly, this type of alignment oversight can be done by agents themselves, not necessarily by humans, but for that, agents need to be aligned with safety protocols.

Risk:
Agentic AI systems acting with full autonomy (e.g., in financial markets, emergency response, or military applications) may not be effectively governed by existing oversight mechanisms, and human intervention may not be practically feasible.

Potential Fix and Policy Recommendations:
Introduce mandatory intervention points or require mechanisms to halt autonomous operations in high-risk applications. Establish risk-tiered human oversight models:

- Full autonomy allowed: Only in low-risk applications (e.g., virtual personal assistants for entertainment or scheduling).
- Supervised autonomy: Requires real-time monitoring for moderate risk (e.g., AI-driven customer service in regulated industries such as finance or healthcare).
- Human-in-the-loop: Required for high-risk AI, e.g., medical diagnostics.

Certification for Fully Autonomous AI

- Introduce an AI Autonomy Certification for any system that operates without human intervention for a prolonged period.
- Regulators could impose usage restrictions if an AI system fails safety audits.

6.2. Risks to Economic Stability and Social Well-being

Beyond its documentation-focused requirements for GPAI, the EU AI Act does not yet fully address the macroeconomic risks posed by Agentic AI (e.g., monopolization, job displacement, or market distortion). These broader, systemic risks may merit consideration as a distinct category not currently covered by the Act. A new category of 'systemic macroeconomic risks' may be needed, while recognizing that the EU Treaty may not provide a clear legal basis for regulating such risks within the existing framework.

Risk:
If Agentic AI, in the form of virtual employees [23], dominates economic decision-making, it could cause disruptions to employment on a large scale, exacerbate inequality, market instability, or power imbalances.

Potential Fix and Policy Recommendations:
This risk may not fall within the scope of the EU AI Act; however, a cross-reference to market regulation and fiscal measures may be necessary, such as updating antitrust rules to consider "functional intent" from autonomous systems or the IMF report [27].

Taxation, Antitrust, and Competition Law Updates for AI:

- In line with "Human-Centric AI", regulation may be needed that promotes social responsibility by creating incentives to retain human labor and by considering tax measures in analogy to the IMF report [27] to limit the replacement of human workers with AI-powered full-time virtual agents, as highlighted in the recent announcement by Anthropic [19].
- Prohibit predatory pricing driven by Agentic AI systems (where AI autonomously lowers prices to eliminate competitors). Enforce transparency in AI-driven pricing and the manipulation of consumer behavior.

Algorithmic impact assessment for market fairness:

- Before deploying an Agentic AI system in areas such as human employment, wages, and market position, an Algorithmic Impact Assessment (AIA) could help understand the risks to employees. The company should consult with employee representatives (e.g., works councils or trade union delegates) to disclose how this system impacts employment, competition, and wage structures. A core challenge is that AI tends to automate discrete tasks rather than eliminate entire jobs, complicating the measurement of employment impact. However, this task-level automation can still have significant consequences: it may degrade job quality, shift skill demands, suppress wages, or reduce working hours. For this reason, an AIA should be designed to capture task-based transformations and their broader market effects.

Reskilling and AI Transition Policies:

- Establish AI reskilling programs to counteract job displacement from Agentic AI.
- Incentivize companies to upskill workers, rather than replace them with autonomous AI.

6.3. Continuous Learning and Unpredictability

The EU AI Act may not encompass the effects of interactions between multiple AI agents. Multi-agent systems have previously been considered in fields such as robotics research; however, they are a new phenomenon in the context of Agentic AI and LLMs.

Risk:
If interaction and cross-influencing Agentic AI systems dynamically change their behavior, ensuring compliance over time becomes challenging. Current risk management approaches may not account for real-time behavioral drift, leading to unforeseen risks.

Potential Fix and Policy Recommendations:
Lifecycle Compliance Monitoring

- Require continuous validation for Agentic AI post-deployment rather than a one-time conformity assessment.
  - Require Agentic AI systems to monitor the governance of the entire system.
  - Mandate periodic audits for AI systems that engage in adaptive learning.

Explainability and Model Change Documentation

- Enforce version control and documentation for Agentic AI systems that evolve over time.
- Extend logging obligations for high-risk AI systems to encompass changes in decision-making patterns, ensuring traceability.

Automated Risk Detection and Reporting

- AI models should have built-in risk detection that alerts regulators if the behavior deviates significantly from their original scope.

6.4. Malicious Use of Agentic AI Systems

The EU AI Act focuses on robustness and bias but lacks specific security measures against adversarial attacks or the misuse of Agentic AI to counter them.

Risk:
Malicious actors could utilize Agentic AI for cyberattacks, fraud, identity theft, or disinformation campaigns.

Potential Fixes and Policy Recommendations:
Implementing mandatory red teaming and adversarial testing for Agentic AI models in high-risk domains.

- Require pre-deployment adversarial testing for Agentic AI in high-risk domains (e.g., financial fraud detection).

AI Cybersecurity Certification

- Establish an EU-wide certification for Agentic AI systems based on their resilience against hacking and adversarial manipulation.
- Require cyber-resilience testing for Agentic AI systems that handle sensitive user data.

Automated Misinformation and Deepfake Detection

- Mandate built-in misinformation detection of Agentic AI models used in public discourse, news generation, or political contexts.
- Enforce traceability mechanisms to ensure Agentic AI-generated content can be verified and attributed.

Agentic AI Risk Management Practices.

- Incentivize risk measurement, management, and mitigation practices for Agentic AI models.
- Promote collaboration between Agentic AI system deployers to foster standard practices in Agentic AI risk management.

7. Conclusion

In this paper, we highlight serious systemic risks to the economy and the citizens of the EU. These risks arise from the use of new and rapidly evolving Agentic AI, which is not yet fully understood. We therefore recommend that the European Commission address this issue and modernize its legislation accordingly. Furthermore, we highlight a normative societal decision that lies beyond the regulatory scope of the EU AI Act.

Acknowledgments

We are grateful for the discussions and feedback from Giacomo Maria Cremonesi (NEC Laboratories Europe), Michael Giardino (Huawei Zurich Research Center), Francisco Medeiros (FM Tech Consult B.V.), Tom Romanoff (ACM Global Policy Director), and Alejandro Saucedo (ACM Europe TPC AI Subcommittee).

References

Note: Several references are to arXiv preprints, which Cornell University hosts. These papers have not undergone peer review yet, but are included here for completeness and readability, reflecting the rapid pace of research in this field.

[1] Horst, H. A., & Miller, D. (2012). *Digital Anthropology*. Routledge. https://www.taylorfrancis.com/books/edit/10.43 24/9781003085201/digital-anthropologyheather-horst-daniel-miller

[2] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues,* 56(1), 81–103 https://doi.org/10.1111/0022-4537.00153

[3] Marquet, Q., Murugavel, S., Charles, X., & Moudni, O. (2025). AI governance for medical chatbots: Designing a multi-agent controller (MAC) for safety. https://framerusercontent.com/assets/HzL9qNSrfuKz83FcUsAmjg7DA.pdf

[4] Knowles, B., Richards, J. T., & Kroeger, F. (2022). The many facets of trust in AI: Formalizing the relation between trust and fairness, accountability, and transparency https://arxiv.org/pdf/2208.00681

[5] Madiega, T. (2023). "*Artificial intelligence liability directive."* Briefing, European Parliamentary Research Service (EPRS).

[6] Yip, M., & Chan, G. K. Y. (2021). Transplanting the concept of digital information fiduciary AI, data, and private law. In The Theory–Practice Interface (Ch. 6). https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=5394&context=sol_research

[7] Ramirez, J. G. C. (2023). *From Autonomy to Accountability: Envisioning AI's Legal Personhood.* https://www.researchgate.net/publication/37890 4514_From_Autonomy_to_Accountability_Envisioning_AI's_Legal_Personhood

[8] European Commission High-Level Expert Group on AI (2019). *Ethics Guidelines for Trustworthy AI* https://digitalstrategy.ec.europa.eu/en/library/ethicsguidelines-trustworthy-ai

[9] Gabriel, I. et al. (2024), *The Ethics of Advanced AI Assistants* https://arxiv.org/abs/2404.16244

[10] Russell, S.J., Norvig, P. (2020). *Artificial Intelligence: A Modern Approach.* (4th ed.)., Pearson.

[11] OWASP (2025) *Threat Modeling Report* https://genai.owasp.org/resource/agentic-ai/threats-and-mitigations/

[12] Shavit, Agarwal, Brundage, and Adler (2023). *Practices for Governing Agentic AI Systems.* OpenAI Retrieved from https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

[13] Sing, A. et al. (2025). *Agentic Retrieval Augmented Generation: A Survey on Agentic RAG* https://arxiv.org/abs/2501.09136

[14] Babaei, G., Giudici, P. and Raffinetti, E. (2025). A rank graduation box for SAFE AI. *Expert Systems with Applications, 259*, 125239. https://doi.org/10.1016/j.eswa.2024.12523

[15] European Commission. (2024). Third Draft of the General Purpose AI Code of Practice https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts

[16] Bengio, Y. et al. (2025) *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?* https://arxiv.org/abs/2502.15657v2

[17] Bengio, Y. et al. (2025). *International AI Safety Report*. https://www.gov.uk/government/publications/international-ai-safety-report-2025

[18] Yu, M., et al. (2025), A Survey on Trustworthy LLM Agents: Threats and Countermeasures. *Communications of the ACM.* https://dl.acm.org/doi/10.1145/3711896.3736561

[19] Axios. (2025, April 22) *Exclusive: Anthropic warns fully AI employees are a year away.* https://www.axios.com/2025/04/22/ai-anthropic-virtual-employees-security

[20] Ballegeer, M., Bogaert, M., & Benoit, D. (2025). Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. *Joint ORBEL-NGB Conference* https://biblio.ugent.be/publication/01JS4JWMWY3J2QWQC0R0N5ZBZ7

[21] Calzarossa, M.C., Giudici, P. and Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity, *Artificial Intelligence Review 58* (150) https://link.springer.com/article/10.1007/s10462-025-11141-w

[22] Hammond, L., et al. (2025). *Multi-Agent Risks from Advanced AI* (Technical report#1). Cooperative AI Foundation. https://arxiv.org/abs/2502.14143v1

[23] Stiefenhofer, P. (2025), *Artificial General Intelligence and the End of Human Employment: The Need to Renegotiate the Social Contract* https://arxiv.org/abs/2502.07050v1

[24] Occhipinti, J., et al. (2024). *In the Shadow of Smith's Invisible Hand: Risks to Economic Stability and Social Wellbeing in the Age of Intelligence* https://arxiv.org/abs/2407.01545v1

[25] Occhipinti, J., et al. (2024). *Recessionary Pressures of Generative AI: A Threat to Wellbeing.* https://arxiv.org/abs/2403.17405v1

[26] Kulveit, J., et al. (2025), *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development.* https://arxiv.org/abs/2501.16946v2

[27] International Monetary Fund (2024), *Broadening the Gains from Generative AI: The Role of Fiscal Policies*. https://www.imf.org/-/media/Files/Publications/SDN/2024/English/SDNEA2024002.ashx

[28] Knowles, B., D'Cruz, J., Richards, J.T., and Varshney, K. R. (2023). "Humble AI." *Communications of the ACM 66,* no. 9: 73-79 https://dl.acm.org/doi/pdf/10.1145/3587035

[29] WIRED Magazine (June 2023). *Humans Aren't Mentally Ready for an AI-Saturated 'Post-Truth World'* https://www.wired.com/story/generative-aideepfakes-disinformation-psychology/

[30] Gefen, D.,et al. (2025). The Importance of Distrust in Trusting Digital Worker Chatbots. *Communications of the ACM.* https://cacm.acm.org/research/the-importance-of-distrust-in-trusting-digital-worker-chatbots/

[31] Cohen, M., et al. (2024). Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). https://dl.acm.org/doi/10.1145/3613905.3650818

[32] Beavins, E. (2025, April). CHAI embarks on post-deployment monitoring for AI as FDA lags. Fierce Healthcare. https://www.fiercehealthcare.com/ai-and-machine-learning/chai-embarks-post-deployment-monitoring-ai