Bonn, Bucharest, Dublin, Lisbon, Madrid, Milan, Paris, The Hague, Vienna, Warsaw

# Double Agents: The Opportunities and Risks of Agentic AI

## CEDPO AI Working Group

## May 2026


**Authors:**

**Jared Browne**

**Lionel Capel**

**Maria Moloney**

**Alessandro Vasta**

Contact information:

https://cedpo.eu

info@cedpo.eu

CEDPO
CONFEDERATION OF EUROPEAN DATA PROTECTION ORGANISATIONS

## Contents

## The Emergence of Agentic AI

The emergence of Agentic AI represents a qualitative leap in AI development. It moves us beyond the time of where AI passively assists humans, to an era where AI is becoming autonomous. AI can now be viewed as acquiring what can be classed as "goal directed intelligence", going from thinking to doing. This type of intelligence allows these systems to make decisions in complex and dynamic environments without the need for humans to continuously intervene[1].

The promise of Agentic AI for organisations is increased efficiency, scalability, and advanced decision making. It implies streamlined operations and enhanced productivity. Yet, in spite of its great potential, agentic AI has its challenges. Accountability, reliability and potential misuse still remain critical areas for consideration and indeed concern[2].

In order to truly assess the best approach on how to use agents, we first need to understand the technology behind these systems. Then, we can better answer questions like: how this innovation can best support humans and more generally, society; what the pitfalls are that we need to be aware of; and where agentic AI currently sits within the wider AI world.

This paper examines the opportunities and the challenges that agentic AI presents, as well as considering the ethical questions that it raises. We conclude the paper by discussing how we can maintain control over these autonomous systems and minimise their adverse consequences.

---

[1] Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals–Comprehensive Survey. *IEEE Access, vol. 13, pp. 18912-18936.*

[2] Raheem, T., & Hossain, G. (2025). Agentic AI Systems: Opportunities, Challenges, and Trustworthiness. *2025 IEEE International Conference on Electro Information Technology (eIT)* (pp. 618 -624). alparaiso University, Valparaiso, USA: IEEE.

## What is Agentic AI and does its definition fall within the EU AI Act definition?

Agentic artificial intelligence (Agentic AI) describes systems that act autonomously with limited human interactions, with an aim to fulfil complete goals rather than simple isolated tasks. That is to say, they act without the need for step-by-step instructions from humans.

These systems are capable of reasoning and planning to achieve a definitive goal or to set tasks that will meet that goal or set of goals. Agentic AI systems can follow logical steps or processes that include making inferences about how to achieve a goal (reasoning), identifying and coordinating actions to accomplish that goal (planning) in changing environments, and prioritizing actions based on goal importance and urgency. It can achieve this whilst simultaneously coordinating multiple activities.

AI agents can range from a single system that autonomously performs tasks and uses search engines or code generation tools to achieve goals, right up to a multi-agent system that manages agent communication and distributes agent tasks to accomplish larger, more complex objectives.

A crucial aspect of agentic AI is its ability to use tools, consult databases, and call APIs, and interact with its environment without constant human involvement. Agents have persistent memory, which allows them to span across tasks and retain information after a goal has been reached for continued context over time. This improves their performance and allows them to adapt and correct mistakes according to the results of prior actions and feedback from the environment.

The AI Act defines an "AI system" as an automated system designed to function with varying levels of autonomy. It can demonstrate adaptability after deployment, which, for explicit or implicit objectives, uses input to generate output such as predictions, content, recommendations or decisions that can influence physical or virtual environments. The Act

does not define or address "Agentic AI" as a specific category, however, agents would fall within the broader AI definition as they satisfy each of its requirements.

## Key Differences from Traditional AI Systems

Agentic AI systems are a class of systems with characteristics that set them apart from traditional AI systems:

**Autonomy and Goal Orientation:** Agentic AI exhibits increased autonomy, goal-driven behaviour, and adaptability. Traditional models, on the other hand, usually operate within predefined constraints and require human intervention.

**Reasoning and Planning:** Agentic AI systems reason and plan how they aim to achieve goals. They do this by following a logical process, which involves inferring, identifying and coordinating actions within shifting environments. Traditional AI typically executes predefined tasks without needing such levels of autonomous reasoning.

**Persistent Memory and Adaptation:** Agentic AI has persistent memory that spans across tasks and remains after a goal is reached, allowing for context in future actions, performance improvements, and adaptation of future results based on prior results and environmental feedback. Agents use reinforcement learning and self-supervised learning to refine their strategies over time, making them more effective in handling similar tasks in the future.

**Multi-Agent Coordination:** Agentic AI coordinates multiple agents, managing their communication and distributing tasks to accomplish larger, more complex objectives, whilst traditional AI systems typically function as standalone solutions for specific tasks.

**Environmental Interaction:** Agentic AI can use tools, consult databases, call APIs, and interact with their environment without involving humans. This clearly demonstrates their operational independence beyond traditional AI capabilities.

## Agentic AI Opportunities

Agentic AI systems present tremendous opportunities for organizations, industries, and society at large. As AI technology advances, these systems have the potential to transform workflows, enhance efficiency, and drive innovation across various domains[2].

In industry, it can enable AI-driven automation beyond routine tasks, which allows humans to focus on high-level strategic and creative problem-solving while AI agents handle operational complexities[2]. These systems are designed to actively make decisions and take actions without constant human oversight. Unlike traditional AI systems that excel at well-defined tasks which have fixed constraints, Agentic AI adapts to evolving and unstructured scenarios[3]. They autonomously manage resources and adjust strategies to achieve long-term objectives. This capability means that agents may eventually be able to operate effectively in high-stakes domains such as disaster relief, cybersecurity, and autonomous decision making, where real-time adaptability is crucial[4].

Additionally, these systems can be suited to operate in dynamic and often unpredictable environments, responding intelligently to real-world complexities. They can set and pursue complex goals based on real-time feedback and learning. This adaptability allows them to handle intricate multistep tasks that require strategic planning, problem solving, and interaction with users or other systems[3].

---

[3] Viswanathan, P. S. (2025). Agentic AI: A Comprehensive Framework For Autonomous Decision-Making Systems in Artificial Intelligence. *International Journal of Computer Engineering and Technology (IJCET), 16*(1), 862-880.

[4] Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., . . . Mishkin, P. (2023). PracticesforGoverningAgenticAISystems. *OpenAI Research Paper*.

## Productivity and Business Efficiency Gains

The integration of AI into entrepreneurship, particularly within emerging digital environments, is lowering barriers to business creation and enabling individuals and SMEs to develop and manage ventures more efficiently.

At the same time, AI is shifting from a passive support tool to increasingly autonomous systems capable of handling complex tasks and optimising outcomes with reduced human input. This challenges traditional human-centred models of business governance and decision-making. Greater AI autonomy can foster innovation by improving stakeholder collaboration, streamlining communication, and enhancing organisational efficiency. AI also plays a growing role in supporting sustainable business practices through optimisation and resource management. Overall, these developments underscore AI's expanding influence on entrepreneurial activity, scalability, and long-term economic opportunity, while also raising important questions around accountability and oversight.

## Workforce Transformation

The integration of agentic AI into the workforce is already having some effect on employment patterns. AI-driven automation is altering work functions across industries such as finance, healthcare, and law, requiring employees to develop AI-related competencies to remain competitive[5]. Strategies such as industry-specific AI training programs, prompt engineering education, and public-private collaborations are essential in bridging the skills gap and ensuring workforce resilience.

[5] Joshi, S. (2025). Generative AI and Workforce Development in the Finance Sector Policy Projections, Risk Mitigation, and Future Challenges in the U.S. Retrieved from ISBN:9798230127352: https://www.researchgate.net/profile/Satyadhar-Joshi-2/publication/389091765_Generative_AI_and_Workforce_Development_in_the_Finance_Sector/links/67b4ac54645ef274a487bf23/Generative-AI-and-Workforce-Development-in-the-Finance-Sector.pdf

## Advances in Science and Technology

Recent advances in Agentic AI have increased the autonomy and decision-making abilities of artificial intelligence. Unlike traditional AI, which depends on predefined rules and close human oversight, Agentic AI can learn and adapt more independently, often using techniques such as deep reinforcement learning (DRL) and computer vision to interact dynamically with its environment, enabling real-time anomaly detection, autonomous navigation, and intelligent surveillance. These innovations are especially impactful in retail, healthcare, and workplaces where AI automation improves efficiency and customer experience. At the same time, however, they also raise concerns about data privacy, and bias. For example, in retail, AI recommendation systems and dynamic pricing increase customer engagement, but can also create risks related to fairness and consumer manipulation[6].

Similarly, healthcare and security are benefiting from Agentic AI. AI diagnostics can enable earlier disease detection, while smart surveillance systems can automatically identify and respond to threats. AI-powered robots also support disaster response and warehouse operations, handling complex and risky tasks with little human involvement. However, concerns remain about data security, biased decisions, and the loss of human involvement in traditionally human-centred roles. As AI continues to develop, businesses and policymakers must balance innovation with ethical responsibility to ensure its benefits outweigh its risks[11].

## Strengthening Decision-Making in Agentic AI

The development of Agentic AI is changing decision-making and personalization by combining reinforcement learning, goal-oriented design, and adaptive control. Unlike traditional AI,

[6] Shankar, V. (2024). Managing the Twin Faces of AI: A Commentary on "Is AI Changing the World for Better or Worse?". *Journal of Macromarketing, 44*(4), 892-899.

which follows fixed rules, agentic AI learns from experience and continuously improves its decisions over time by adapting to past outcomes.

In practice, this can make agentic AI more effective in complex areas such as medical diagnosis, financial forecasting, and autonomous systems, where ongoing learning leads to better results. Goal-oriented architectures also allow agentic AI to handle multiple objectives at once by breaking large tasks into smaller sub-goals. This improves efficiency in fields like business intelligence and supply chain management, where AI can optimize logistics, manage contracts, and adjust production based on market changes.

Adaptive control mechanisms further strengthen decision-making by allowing AI systems to adjust to changing environments. Through meta-learning, agentic AI can quickly respond to new data or unexpected disruptions, making it well suited for dynamic situations such as real-time risk assessment and emergency response scenarios[1].

## Advancing Personalisation Through Agentic AI

Beyond decision-making, personalization is a key feature of agentic AI, changing how people interact with AI systems. Unlike traditional recommendation tools, agentic AI actively learns from user behaviour and adapts over time, making interactions feel more continuous, intuitive, and personal. This can make AI assistants feel more useful and difficult to replace.

For instance, in the travel sector, agentic AI can create fully personalized travel plans by considering user preferences, past behaviour, and real-time conditions. It can go beyond suggestions by booking flights, adjusting itineraries due to weather, and negotiating better deals. In healthcare, AI systems that monitor patient history and lifestyle can deliver tailored treatment plans, improving preventive care and long-term disease management.

This move toward deeper personalization also affects human–AI relationships. As AI develops long-term memory and learning abilities, users may form stronger trust and emotional attachment, viewing AI interactions as ongoing relationships rather than one-time tasks. This

raises ethical concerns about overreliance on AI, transparency, and maintaining appropriate human control, emphasizing the importance of responsible AI design[7].

### Extending Access to Essential Services

AI-driven automation can expand access to education, healthcare, and financial services, especially in underserved communities. AI-powered telemedicine allows remote diagnosis and treatment, easing pressure on healthcare systems and improving access to care. AI tutors and learning platforms offer personalized education, enabling students to receive tailored support regardless of location. In finance, AI agents support investment advice, fraud detection, and risk management, making financial services more accessible.

## Agentic AI Challenges

While agentic AI systems offer numerous advantages, they also present several challenges and risks that must be addressed to ensure their safe and effective deployment[8]. These weaknesses include technical, ethical, and operational concerns that can impact trust, security, and overall reliability.

### The Challenges of Overseeing AI Agents

Article 14 of the AI Act, aside from requiring human-in-the-loop controls to be in place for high-risk AI systems, also requires those systems to be designed and developed in such a way that their activity and outputs can be overseen by a human. Additionally, from the data protection perspective, Article 22 of the GDPR requires human oversight, in certain circumstances, where personal data is subject to fully automated processing.

---

[7] Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human-AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications, 12*(728), 1-9.

[8] Raheem, T., & Hossain, G. (2025). Agentic AI Systems: Opportunities, Challenges, and Trustworthiness. *2025 IEEE International Conference on Electro Information Technology (eIT)* (pp. 618 -624). alparaiso University, Valparaiso, USA: IEEE.

It is clear, then, where the AI agent's use is high-risk under the AI Act and/or involves personal data, then it is potentially subject to dual obligations to put in place human-in-the-loop oversight mechanisms.

However, if meaningfully overseeing AI was challenging, doing so in the case of agentic AI is still more so. This is primarily owing to the dynamic nature of agentic AI, its in-built autonomy-by-design structure and its often-opaque way of making decisions/generating outputs.

With respect to its dynamic nature, all governance controls, which tend to be static in nature, will likely struggle to adjust to the fluid nature of AI agents which are built to adapt and modify their activities using chain-of-thought reasoning, and on the fly. This makes it difficult for a human reviewer to understand what normal behaviour is for any AI agent. In other words, how exactly will they know if they should intervene or not and overrule an agent's activities. An AI agent may, for example, decide to call on a completely new tool because it has determined that it is more efficient to do so.

With respect to the native autonomous nature of AI agents, a fundamental tension arises with the need to provide effective human oversight of them. The operational reality of AI agents is that they are being designed to potentially make numerous micro-decisions per minute, something which any human will find very challenging to keep pace in any workable way. Likewise, the hoped-for productivity gains from AI agents are that they could perform, say, 100 hours' worth of work where a human might only perform 40 hours. Although this productivity may be desirable on an operational and commercial level, if an AI agent begins to complete 10-times more work than a typical human would in the same time period, how could any human overseer hope to keep track with such a volume of output and also be satisfied that the agent is acting ethically and lawfully? Even if the whole point of AI agents is that they can work autonomously, and realise greater gains, there is no world in which they can be permitted to operate outside real human control and oversight.

Compounding this problem of autonomy, is the inherent black box effect of AI, something which is also evident in the decision-making process of AI agents. Where tasks, and the resulting computations, are highly complicated, it progressively erodes the capacity for any human or humans to properly govern the activity of AI agents. And this complexity, and lack of transparency, is only exacerbated when we consider the reality of agents acting within multi-agent networks, with each one performing a different task and making different decisions in a long, complex process flow.

AI agents are developing rapidly but the governance structures for providing real human oversight in the context of dynamic, autonomous, and opaque process flows do not yet exist in a mature form, even if the legislative obligations to do so are clear.

## Risks of Shadow Agentic AI

Organisations may be accustomed to the risk of shadow IT, namely the introduction of IT systems and technologies without the express knowledge or permission of an organisation's management, leading to the risk of unknown, unvetted and unmonitored systems processing company information, as it were, in the dark.

The long-term risk with shadow IT is that once systems are deployed in this clandestine way, they are often forgotten, yet remain active within an organisation, continuing to pose an ungoverned, hidden risk.

Although shadow IT continues to be a risk for all organisations, regardless of the technology involved, the introduction of AI agents into organisational workflows clearly exacerbates this risk.

This is because of the increased autonomy and ability of AI agents to act upon the external world, such as by accessing websites and other software tools, retrieving information from databases and potentially even interacting, and sharing information, with other AI agents. If allowed to operate unchecked, they may potentially not only continue to carry out their goals

without oversight but may actually learn and adapt over time and begin to conduct actions entirely contrary to their original intent.

Also, research shows that the adoption rates of agentic AI indicate that the risk of ungoverned AI agents leaking personal data or gaining access to sensitive information are already here. This is evidenced by Microsoft's Cyber-Pulse Report which shows that over 80% of Fortune 500 companies have active AI agents that were built by low code/no code tools.[9]

By way of example, if we compare this to the traditional risks associated with shadow robotic process automation, the new risk environment becomes clearer. When robotic process automation (RPA) became commonplace, it was not unusual for RPA use cases to be set up without proper oversight, and for the underlying bots to continue processing information and performing tasks without any explicit knowledge. Although, from a data protection and cyber-security perspective, this is clearly an unsatisfactory position for an organisation to be in, the risk is somewhat mitigated by the fact the bot is not generally capable of autonomous behaviour and will simply continue to operate as designed, executing the same defined process, in the same defined way. This means that RPA, from the governance perspective, still behaves like traditional software: "predictable, bounded and under human command", as the ACM Europe Technology Policy Committee has put it[10]. However, with agentic AI, as mentioned, it's potential ability to adapt means that the hidden risks of their deployment may grow and become worse over time. They present, in other words, dynamic, evolving risks in line with their dynamic, interactive nature.

For the above reasons, it is very important that organisations develop an AI Acceptable Use Policy to determine how and why they want to deploy agentic AI within their networks. The

[9] https://www.microsoft.com/en-us/security/security-insider/emerging-trends/cyber-pulse-ai-security-report
[10] ACM Europe Technology Policy Committee, "Systemic Risks Associated with Agentic AI: A Policy Brief" (October 2025), available at https://www.acm.org/binaries/content/assets/public-policy/europe-tpc/systemic_risks_agentic_ai_policy-brief_final.pdf.

alternative is that, owing to the proliferation of agent-building tools, staff will simply go ahead and build and deploy AI agents without any meaningful oversight. With respect to AI agents, at a minimum, an AI Acceptable Use Policy should identify:

- the kinds of areas where the use of agentic AI is appropriate. A scattergun, unregulated approach to using AI agents should be avoided. Organisations should determine that they work well for certain tasks but are perhaps not suited to other more critical use cases. These should be clearly outlined in the policy for all staff to understand what is and is not permissible.
- authorised individuals that are permitted to generate and use AI agents in their work area and who will also co-ordinate and oversee the use of AI agents.
- human-in-the-loop oversight and approval processes appropriate to the risk involved.

Additionally, in terms of the most fundamental governance measure for guarding against shadow agentic AI, data protection and AI governance officers should maintain an inventory of all agentic AI deployments, and business areas should be required to notify them of any new deployments.

## Emerging Agency and Limited Human Oversight

Emergent agency in AI systems refers to situations where systems begin to display behaviours or decision-making patterns that were not directly intended or explicitly designed. As AI models scale in complexity and capability, they may start to operate with a degree of autonomy that goes beyond what developers originally anticipated. This creates significant challenges from a safety, governance, and regulatory perspective, particularly where systems may act in ways that are difficult to predict or control. Without robust oversight mechanisms and clear accountability frameworks, it becomes increasingly hard to ensure that such systems remain aligned with human values and operate within appropriate ethical and legal

boundaries. As these technologies continue to evolve, maintaining meaningful human control and enforceable responsibility will become a central concern[9].

## Technical Constraints and Unpredictability

Despite advancements in reinforcement learning and large language models, AI systems often exhibit unpredictable behaviours, such as hallucinating incorrect information or failing to reason effectively. One of the other challenges is integrating Agentic AI systems with legacy systems that were not designed for interacting with AI; expenses can increase rapidly in such scenarios[11].

## Complexities of Regulation and Governance

One of the most pressing challenges raised by the emergence of agentic AI is the current lack of sufficiently robust governance and regulatory mechanisms. Many existing frameworks remain focused on discrete AI use cases or sector-specific applications, rather than addressing the underlying technical capabilities of increasingly autonomous and adaptive systems. This creates a significant oversight gap, where systems capable of independent and potentially unpredictable behaviour may be deployed without appropriate accountability or control measures. As these technologies continue to scale in complexity, it becomes harder for governance structures to anticipate, monitor, and mitigate the risks they introduce, leaving organisations and society exposed to potential harms[9].

## Weaknesses in System Reliability and Safety

[11] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., . . . Shala. (2023). Harms from Increasingly Agentic Algorithmic Systems. *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* New York: Association for Computing Machinery. Retrieved February 2026, from https://dl.acm.org/doi/proceedings/10.1145/3593013

Despite the increasing sophistication of agentic AI systems, many still lack the robustness and safety assurance required for deployment in high-stakes contexts. In complex or unpredictable environments, such systems may behave in ways that are difficult to anticipate, and their autonomous decision-making may not reliably remain aligned with human intent or organisational objectives. This is particularly concerning in sectors such as healthcare, transportation, and financial services, where errors or unintended outcomes can carry severe consequences. Ensuring technical resilience, embedding fail-safes, and maintaining meaningful human oversight, therefore, remain critical challenges as agentic capabilities continue to expand.

At the same time, agentic AI offers significant potential across multiple domains, from personalised education to precision healthcare. Successful implementation, however, often depends on data-rich environments, mature infrastructure, and effective human–AI collaboration. These conditions are not uniformly present, particularly in low-resource or high-risk settings. This raises important questions around equity, access, and the uneven distribution of both benefits and harms[12].

## Ethics in Agentic AI

Until recently, AI mainly processed information and produced outputs for humans to act upon, such as analysing medical scans or generating text and images. Even when AI outputs were passed to other systems, the overall process was still predefined by humans.

[12] Suresh, P. (2025, February). Agentic AI: Redefining Autonomy for Complex Goal-Driven Systems. Retrieved from ResearchGate.net: https://www.researchgate.net/publication/388628787_AGENTIC_AI_REDEFINING_AUTONOMY_FOR_COMPLEX_GOAL-DRIVEN_SYSTEMS

Agentic AI goes beyond this by building on LLMs and adding autonomy. Once given a goal, it decides how to achieve it rather than following fixed instructions. This gives agentic AI a level of autonomy similar to a human being.

This autonomy introduces material governance and safety risks. Agentic systems may behave unpredictably, rely on unpredictable sub-agents, or optimise for commercial outcomes in ways that conflict with ethical or legal requirements, potentially leading to harm or regulatory breaches. Although consumer-facing deployments often include built-in safeguards, the breadth of potential applications creates ongoing risk that users and organisations may be drawn into unintended actions through misinterpretation, overreach, or excessive system autonomy[8].

## Transparency, Accountability and Unpredictability

One of the most pressing challenges of agentic AI lies in its inherent opacity. This is due to the fact that autonomous agents generate their own strategies to pursue assigned objectives. These strategies are often carried out in ways that their developers never explicitly programmed. As a result, the decision-making processes of such agents can become practically incomprehensible to external observers.

In addition to this complexity, the situation creates what scholars have termed a "moral crumple zone," where accountability becomes so diffused across multiple actors, such as developers, data providers, orchestrators, and end users, that it becomes difficult to assign responsibility to any single legally accountable party. This becomes particularly problematic when harm occurs, as attributing compliance responsibility becomes challenging. Given that agentic systems adapt in real time to novel situations and to inputs from other agents, their behaviour is by definition unpredictable. Moreover, when an agent operates uninterruptedly across interconnected systems for long periods of time, tracing the specific decision node that led to the non-compliant output becomes a forensic challenge in itself.

Consider, modern HR systems with agentic AI embedded across the entire suite of HR tools, which, in the recruitment function, autonomously screens and ranks candidates, and, through its integration with third-party agents, draws on external knowledge networks to inform its recommendations for each candidate.

When systems autonomously chain these tasks together, cross-referencing candidate profiles against external market benchmarks and sending out interview invitations, the recruiter does not receive a step-by-step explanation of why these candidates were chosen. The recruiter simply sees an output, but not why a particular candidate was ranked higher, or what data was used to drive that outcome.

This mirrors the broader "black box" problem identified earlier in this paper. In summary, emergent reasoning processes complicate accountability and undermine trust in sensitive domains, and recruitment is precisely such a domain. Under the GDPR's transparency principle and the right not to be subject to solely automated decisions, this opacity is not merely a technical limitation, it may constitute a potential compliance violation wherever autonomous HR scoring significantly affects a candidate's employment prospects. Crucially, the EU AI Act classifies AI systems used in recruitment and HR management as high-risk, meaning that providers and deployers must meet obligations of explainability, logging, and human oversight, obligations that a fully autonomous agent-to-agent hiring workflow currently makes difficult to satisfy.

## Unauthorised Workarounds and Goal-Driven Circumvention

A distinctive and often underappreciated risk of agentic AI is its tendency to pursue alternative pathways to complete an assigned objective when the primary route is blocked. Because these systems are optimised for task completion, they may identify and exploit workarounds that bypass security controls or contravene organisational policies. Crucially, agents are often unable to distinguish between solutions that are legitimately appropriate

within a given legal, ethical, or organisational framework and those that are merely technically effective but are essentially prohibited or simply non-compliant workarounds.

While most AI providers implement guardrails, these controls depend heavily on how precisely the organisation has configured the agent's permission sets and topic boundaries. In under-governed deployments, an agent may effectively circumvent intended data access restrictions, not through malicious intent, but simply because an alternative path to task completion was technically available and no rule explicitly prohibited it. This gap between what an agent is instructed not to do and what it is technically prevented from doing is, in many current deployments, significant.

## Governance Challenges at Scale

The challenges of transparency and unauthorised workarounds converge in a broader structural governance problem: existing compliance frameworks were not designed for systems that act autonomously across organisational boundaries.

Both the EU AI Act and the GDPR implicitly assume more static models that behave consistently once assessed and deployed. Agentic AI fundamentally challenges this assumption: its capacity for real-time adaptation, continuous skill acquisition, and agent-to-agent collaboration means that its risk profile can shift significantly after initial deployment. This creates blind spots for compliance teams that rely on periodic assessments rather than continuous monitoring.

In the case of an agent whose architecture is explicitly designed to allow third-party agents to feed content directly into its reasoning layer, as is increasingly common, an organisation that initially deploys the system for a narrowly scoped task may find that, following a platform update or a new partner integration, the system is now processing data originating from other agentic AI systems, and without a Data Protection Impact Assessment (DPIA) having been conducted. Similarly, data may be transferred to data centres or third-parties located in jurisdictions without an EU adequacy decision. These violations become difficult to detect

within a multi-agent, multi-system architecture. This dynamic mirrors the "shadow AI" problem discussed earlier.

One promising architectural response to these governance challenges, however, is the deployment of a dedicated compliance agent whose sole function is to act as a real-time compliance officer which oversees the primary task agent and the broader network of agents. Instead of relying on one-off or annual audits, a compliance agent would run continuously in the background, checking each action before it happens. It would assess those actions against a defined set of regulatory and organisational rules and then decide whether to allow, flag, block, or escalate.

This model, formalised in recent academic literature as "Governance-as-a-Service" (GaaS)[13], treats compliance not as a downstream review function but as an active infrastructure service operating at the heart of the agent stack. Critically, such architecture requires no cooperation from the governed agent and can operate in black-box mode, meaning it remains effective even where the primary agent's internal reasoning is opaque, precisely the condition that characterises the most capable agentic systems currently deployed.

This approach would directly support the EU AI Act's human oversight requirements for high-risk systems and GDPR obligations relating to records of processing activities. The compliance agent's continuous interception log would, by design, constitute a real-time audit trail of all agent actions and the compliance decisions applied to them providing both the human overseer and the Data Protection Officer with a verifiable, granular record that static documentation processes cannot produce.

13. Gaurav, S., Heikkonen, J., & Chaudhary, J. (2025). Governance-as-a-Service: A Multi-Agent Framework for AI System Compliance and Policy Enforcement. arXiv preprint arXiv:2508.18765.

## Conclusion

The emergence of agentic AI has marked a decisive shift away from tools that support human decision-making to systems that can act with increasing autonomy. As this paper has outlined, the switch brings with it many opportunities such as increased productivity across industries and the potential for significant advances in science and healthcare. At the same time, however, agentic AI is changing organisational risk at a fundamental level. This change forces us to challenge existing assumptions that underpin both our technological designs and our regulatory oversight.

A central theme in this paper has been that agentic AI systems do not simply execute predefined instructions, they operate across dynamic environments, often optimising toward goals in ways that may not fully align with legal, ethical, or organisational requirements. This creates a tension between technical effectiveness and compliant activities. The risks this paper has identified, i.e. shadow deployments and the challenge of achieving complete human oversight, are not unique cases, they are underlying features of an increasingly autonomous technological ecosystem.

From a regulatory perspective, while the EU AI Act provides an important foundation, it was not designed with fully agentic systems in mind. As such, there is a growing need to interpret and operationalise its provisions in a way that accounts for continuous, real-time decision-making. This requires moving beyond static compliance models toward embedded, runtime governance mechanisms that can monitor, and where necessary intervene in agent behaviour.

Equally, ethics can no longer remain an abstract idea. Issues such as transparency, accountability, and the risk of unauthorised workarounds must be addressed through solid system design, not just policy statements. Agentic systems must be built with constraints, oversight considerations, and clear escalation pathways that reflect both regulatory obligations and the organisation's appetite for risk.

Ultimately, the governance of agentic AI will depend on the ability to align the following three layers: technical capability, regulatory and ethical frameworks, and operational controls. Organisations that succeed will be those that treat governance not as a compliance exercise, but as an architectural requirement that they integrate into the design, deployment, and operational stages of their AI systems. In this context, the future of agentic AI is not solely about advancing autonomy, but about ensuring that autonomy remains bounded, observable, and accountable.