ICO tech futures: Agentic AI

1.0.0 | 8 January 2026

ICO tech futures: Agentic AI

Foreword

In this Tech Futures report on agentic AI, we set out our understanding of the emerging technology, including its potential uses and expected technical developments. We share our early thoughts about the data protection implications that organisations will have to consider as they explore the deployment of agentic AI, including data protection risks and opportunities. We share four possible scenarios to explore the uncertainty around how organisations might adopt agentic AI and how its capabilities might develop over the next two to five years.

Executive summary

Agentic artificial intelligence (AI) is evolving at pace, attracting intense scrutiny from innovators, technology adopters and regulators worldwide. As organisations consider deploying agentic AI, understanding its capabilities and the associated risks is essential.

Agentic AI combines the capabilities of generative AI with additional tools and new ways of interacting with the world. This increases the ability of AI systems to work with contextual information, operate using natural human language and automate more open-ended tasks. Agentic AI systems are being developed for use in research, coding, planning and transactions. Their potential applications span commerce, government, the workplace, cybersecurity, medicine and the consumer space. Many believe that agentic capabilities can form the foundation for powerful personal assistants.

While agentic AI offers some new technological capabilities we are at an early stage in development, with many use cases unproven or at the development stage. At the ICO, we are building a well-informed evidence base about:

- where the technology is now; and
- how to exercise caution about the proven abilities of agentic AI while identifying and managing the data protection issues and risks related to supporting privacy-led innovation.

As developing agentic AI increases the potential for automation, organisations remain responsible for data protection compliance of the agentic AI they develop, deploy or integrate in their systems and processes.

We have already explored in our consultation series on generative AI the many issues that agentic AI shares. **Novel agentic AI data protection risks** include:

- issues around determining controller and processor responsibilities through the agentic AI supply chain;
- rapid automation of increasingly complex tasks resulting in a larger amount of automated decision-making;
- purposes for agentic processing of personal information being set too broadly to allow for open-ended tasks and general-purpose agents;
- agentic systems processing personal information beyond what is necessary to achieve instructions or aims;
- potential unintended use or inference of special category data;
- increased complexity impacting transparency and the ease with which people can exercise their information rights;
- new threats to cyber security resulting from the nature of agentic AI; and
- concentration of personal information facilitating personal assistant agents.

One of our key findings from this initial work is that the specific design and architecture of agentic systems impact how data protection law applies and how people exercise their data protection rights. Choices such as the data and tools that a system can access and which governance and control measures to put in place really matter.

Poorly implemented agentic systems will increase the risks of data protection harms. For example, this could include systems that:

- have no clear purposes;
- are connected to databases not needed for their tasks; or
- have no measures in place to secure access, monitor or stop activity, or control the further sharing of information.

The importance of design and architecture also means that there are good opportunities for privacy by design and privacy-friendly innovation in agentic AI, and organisations should use them for responsible deployment. We are already seeing some features and tools intended to address privacy issues.

We have identified **innovation opportunities** with agentic AI that have the potential to support data protection and information rights and contribute to privacy-positive outcomes. Potential areas include:

- data protection compliant agents;
- agentic controls;
- privacy management agents;
- information governance agents; and
- ways to benchmark and evaluate agentic systems.

Due to the pace of development of agentic AI and the speed at which developers are experimenting, we are trying two new approaches with this report. We are using **scenarios** of four different potential futures to explore the uncertainty about how agentic AI might be adopted and how its capabilities might develop over the next two to five years.

The ICO's role

Our aim at the Information Commissioner's Office (ICO) is to ensure that innovation in agentic AI develops in ways that protect people's information rights, while providing clarity and support for organisations. Our next steps on agentic AI include the following:

- Hosting workshops with industry to gather further information on agentic AI, including on agentic capabilities and adoption, and how industry is mitigating data protection and privacy risks.
- Updating guidance on automated decision-making and profiling, in light of the Data (Use and Access) Act, starting with public consultations in 2026.
- Working with partner regulators through the Digital Regulation Cooperation Forum (DRCF) to understand the cross-regulatory implications of agentic AI and invite innovators to participate in the Thematic Innovation Hub on agentic AI.
- Continuing our work with international partners through the G7 Data Protection Authorities Emerging Technologies Working Group.
- Inviting stakeholders working on agentic AI applications to access our innovation support services. For organisations that are in the process of developing innovative products and services using personal information and agentic AI in the public interest, we encourage them to explore our Regulatory Sandbox.

We would like to encourage and support data protection–focused opportunities in agentic AI. We will address innovation opportunities proactively as agentic AI matures and our role in regulating it develops.

We will keep our approach under review as technologies, markets and risks evolve.

Introduction

Tech Futures is our technology foresight series. Each report explores an emerging technology, and we share our early thinking and understanding. The reports are not guidance or formal regulatory expectations but part of our process for responding to new technologies. In this report on agentic AI, we:

- explain our understanding of agentic AI;
- identify its potential developments and use cases;
- highlight potential data protection issues that may emerge with increased adoption of agentic AI; and
- present potential innovation opportunities that could support information rights.

We have developed the report based on desk research, stakeholder interviews and futures methodologies. Annex 1 provides more detail on the methodology.

Why agentic AI?

In June 2025, we published Preventing harm, promoting trust: our AI and biometrics strategy. In this strategy, we committed to:

- work with industry to explore the data protection implications of agentic AI over the next two to five years; and
- publish a Tech Futures report addressing issues such as accountability and redress over the longer term.

Our ambition is to encourage responsible development and use of agentic AI.

There is a long-term ambition in the field of AI for sophisticated AI assistants and automating complex activities. Recent technological advances open up the possibility of AI agents with increased capabilities.

Prominent technology companies are talking about their intention to develop or use AI agents. [1] Some consider that agentic AI offers the potential economic pay-off for investment in generative AI over recent years.[2]

As such, there have been a range of recent announcements, including:

- updates to some major LLMs, including updates designed to enable more complex agents [3] or functions such as agentic web browsing; [4]
- release of dedicated tools for building agents using these models; [5]
- anticipated release of an experimental 'agentic workspace' that will enable developers to test using agents to complete tasks on their computer; [6]and
- release of an agentic AI platform by a major customer relations platform provider. [7]

Reports indicate that the spread and deployment of AI is having a large impact on the UK's economy. AI companies raised over £14 billion in revenue in 2023. [8] The number of AI companies is growing, with a 2024 study showing a 17% increase from the previous year. This growth could continue as agentic AI continues to develop and organisations adopt it more widely. The UK Government has suggested that agentic AI could help to revolutionise the way people interact with public services. [9]

Some predict that the spread of AI and the eventual deployment of agentic AI could have a bigger impact on the world economy and finance than the internet. Others are more cautious, expressing concern that some commentators exaggerate agentic AI's potential and capabilities.

We consider that the technical advances, combined with the market attention, require us to understand potential developments in this area. We must be able to separate hype from real potential.

The widespread use of AI agents could raise challenges for privacy and data protection, including accountability, transparency, data minimisation and purpose limitation. If organisations fail to demonstrate adequate compliance of their AI agents, this could risk undermining public trust. Without that trust, people may be less willing to support or work with AI-powered services. This creates a barrier to responsible adoption across the UK economy.

There are also potential opportunities for innovations in agentic AI that could support data protection, privacy and information rights. For example, organisations can use privacy by design in creating agentic systems or agents could support or automate the exercising of information rights. As supporting responsible innovation is part of our contribution to economic growth, we would like to encourage these opportunities by drawing attention to them.

This report accompanies internal work we are doing to understand how agentic AI is developed and used in the real world, as we develop a statutory code of practice on AI and automated decision-making. It builds upon our previous Tech Horizons report on Personalised AI and our published guidance on AI and data protection. Additionally, we have published the outcomes of our consultation on generative AI, which sets out our analysis of how specific areas of data protection law apply to generative AI systems.

This report explores different regulatory scenarios so we can be ready for the various ways agents might be adopted.

What are agentic AI and AI agents?

Definitions of agentic AI and AI agents vary. Organisations use a variety of terms to market and promote these technologies, and these can differ from definitions in scientific literature. In this section, we set out how we understand agentic AI for the purposes and scope of this report.

In computing, an **agent** is software or a system that can carry out processes or tasks with varying levels of sophistication and automation. One relevant example is automatic stock management and ordering. Previously, agents have typically been specialised and designed to perform specific tasks within pre-set limits.

Recently, advances and new approaches in AI have increased the potential autonomy and range of tasks that people may give AI agents. This is leading to novel applications and tools with new capabilities.

Large language models (LLMs) are statistical models trained on vast amounts of language. Along with other types of **foundation models**, they are one of the technologies behind the generative AI tools released in recent years. When LLMs or foundation models are integrated ('scaffolded') with other tools, including databases, memory, computer operating systems and ways of interacting with the world, they create what industry is calling **agentic AI**. This specific form might also be called 'agentic LLMs'.

An agentic system is any computing system that makes use of this agentic capability. The agentic nature of a foundational model can vary significantly depending on which tools it is scaffolded with.

This approach has fundamental differences from previous agent systems. LLMs and other highly capable general-purpose AI models can enable agentic systems to:

- work with contextual information;
- take instructions and provide information in natural language;
- use knowledge embedded in their training data;
- handle various types of information; and, potentially,
- perform iterative 'reasoning'.

They can function in a wider range of circumstances and potentially allow for long-term planning, goal-oriented behaviour and adaptive decision-making.

Importantly, developers are designing modern AI agents that can create and execute context-specific plans in more variable environments, with less human direction. This means that, while traditional software typically follows a fixed way to solve problems, agentic AI might generate different ways to approach a problem or achieve a goal. With appropriate scaffolding, these systems could have the tools to put those plans into action and affect the real world. However, because they build on LLMs, some of the negative characteristic features of LLMs may be present, such as:

- making up facts ('hallucinations');
- providing confidently expressed but incorrect answers; or
- expressing bias embedded in their training data.

The novel nature of agentic systems may result in unanticipated or unprecedented actions as they work towards completing their goals. Understanding the reasoning behind their actions may require an intensive investigation of logs and error conditions. This particularly applies where an AI may have hallucinated to reach a conclusion or given wrong answers presented as fact.

Agentic AI systems are likely to display, to some extent, the following four capabilities:

- Perception – being able to work with a wide range of potential inputs. This could include natural language and unstructured sources created for different purposes that were not designed to be machine-readable.
- Planning or reasoning-like actions – for example, generating plans, dividing tasks into sub-tasks and checking for errors. [10]

- Action – including accessing tools, interacting with people or other agents and generating and running code.
- Learning and memory – adaptive decision-making, incorporating error corrections into future plans, learning the preferences of their users and learning from feedback.

Agentic AI systems may take different forms depending on their intended use. 'AI assistants' are often mentioned as a potential use of agentic AI. However, beyond using agentic AI as a standalone agent, organisations can also use it:

- in background processes;
- as part of widespread infrastructure;
- in operating systems; or
- as part of other software.

In addition, several agents can be combined to form '**multi-agent systems**'.

Researchers have started to evaluate agentic AI capabilities on a range of measures: [11]

- Autonomy – How independently the AI system can operate, or how little human input is needed either to oversee a task or make it work correctly.
- Efficacy – The ability to interact with and impact the world.
- Goal complexity – The complexity of goals that the agent can handle. These can range from simple goals with simple solutions to complex competing goals that require a series of subtasks to achieve. A component of this is long-term planning – the extent to which the agent can work on goals over longer periods of time.
- Generality – The ability of an agent to work across different roles, contexts and types of tasks.
- Underspecification – The extent to which an agent can achieve a goal without someone specifying how it should do so. [12]

In this report, our focus is on the emergence of AI systems that have limits but can:

- autonomously pursue goals;
- adapt to new situations and contexts; and
- exhibit some reasoning-like capacities.

We are not concerned with agents that are far more limited, task-specific and cannot learn and implement new approaches. Agents such as these are already on the market and outside the scope of this report. For example, we have observed the term 'AI agents' used to describe chatbots that are based on LLMs but not integrated with external tools.

We are also not addressing the potential emergence of artificial general intelligence (AGI), which refers to hypothetical AI systems that can match or exceed human-level performance across all tasks, or any associated existential risks.

Potential use cases

Our research and engagement with stakeholders raised several situations where organisations already use AI agents. Current day and near-future uses include:

Research – Researchers using agentic systems to assemble reports or summaries based on information sources they have access to. Examples include producing sales reports and analysis or helping with scientific research.

Coding – Programmers are already using agents to support coding. They can develop new scripts or check the output of human coders in real-time.

Planning, organising and executing transactions – Current and near-term emerging functions include basic planning tasks, such as:

- booking a meeting while taking into account available meeting times;
- planning and booking travel by searching the web; and
- asking a user to pick from recommended purchase options.

We also found current examples of agents able to automate data entry or complete more advanced customer service tasks.

More advanced, near-term versions might:

- manage your diary engagements more proactively; and
- plan and book social activities from prompts you give them.

Further into the future, agents may interact autonomously with various third-party agents. This may lead to use cases based on marketplaces or ecosystems of multiple agents.

Potential use cases further along our timescale are less clear. However, some AI developers are encouraging and supporting experimentation with novel use cases in the workplace.

Some stakeholders suggested that, in future, agentic AI would be able to do 'anything' and people and organisations could widely deploy it throughout systems and workflows. Other stakeholders suggested that UK deployments in high-risk contexts where errors could incur legal liability (such as law, finance or medicine) may happen more slowly.

Stakeholders saw potential uses across many industries in the future as agent capabilities and user comfort increase. We could see wider integration with different tools, greater interoperability between agents and improvements in early use cases. Examples could include the following:

Agentic commerce

We are already seeing early agentic commerce agents. [13] For example, early systems use agents to present a customer with a pre-populated list of retailers for something they want to buy. They then search the web for options, taking into account budget and personal preferences set out in a prompt. Once they've made a selection, and with appropriate permissions, the agents could arrange payment and coordinate with the customer's diary to arrange delivery.

Organisations are already integrating more advanced agents into customer service roles. For example, an agent may access multiple different customer management systems to check if a received complaint is valid. In future, they could automate tasks (such as refunding transactions, filling out forms or booking appointments). Organisations could also use agents to automate promotions by generating hyper-personalised campaigns and interacting with customers as needed across multiple communication platforms. [14]

It's possible that long-term, agentic commerce could become standard, with customers and companies delegating more transactions to their agents to handle directly. We may see customers' personal agents anticipating shopping needs and making proactive purchases. These could be based on broad objectives, learned and defined preferences or behaviours, or knowledge of upcoming plans, rather than specific prompts.[15]

For example, if someone were planning a large purchase, their shopping agent may connect with their bank account or financial agent to check if the

purchase is within their budget for the month. The shopping agent may conduct market research and schedule large purchases to get the most benefit from sales and discounts or negotiate a price directly with a seller. The financial agent may then adjust future planned spending to allow for the purchase and provide details of how this affects other spending plans. This could potentially extend to agents seeking out tailored financing options to present to the shopper for agreement.

Workplace applications

Agentic systems could improve individual and organisation functions. This might include using AI assistance to create smaller, more agile teams – or even having some functions or specialisms entirely handled by an agent.

An example of such an 'agentic workflow' could be an agentic IT assistant. If an employee contacts the agent with an IT issue, the agent could ask clarifying questions, take steps to diagnose and resolve the IT issue, and learn from the result. [16] This approach could also apply to other office functions, such as recruitment. Early agents could automatically generate a job description and filter applicants, conduct chatbot-based first-round interviews, book additional interviews and provide feedback. [17]

Or in the case of insurance, agentic AI could automate data entry, review unstructured claim records, flag potential evidence of fraud and generate recommendations. [18] One stakeholder highlighted the potential for workplaces to create 'digital twins' of a person's job. We are already seeing early examples of this. [19]

Government services

The UK Government is exploring the possibilities of deploying agents in some government services by 2027. Early experiments will focus on using agents to provide tailored employment guidance and support. The future intention is to experiment with agentic systems that could automate some government 'life admin'. This could include, for instance, automatically updating addresses and electoral registration or signing up for a new GP when a person moves house. [20]

Other examples could include helping social services users complete administrative tasks and access online services. This would give social services professionals more face-to-face time with service users. [21]

Automated cybersecurity applications

People and organisations are likely to use agents in future to both secure systems and attack them. Increasingly autonomous agents could enable wide-scale attacks on systems with little human involvement. [22] In the near term, they could scale up human attacker capabilities. People could task agentic systems to probe, examine and create custom attacks on remote networks.

However, agentic systems also offer opportunities for more advanced automated defence systems by protecting networks from cyber threats. Agentic systems could complement existing methods of detecting malicious activity by:

- proactively identifying vulnerabilities before they are exploited; or
- acting reactively to safeguard networks.

In both circumstances, they could either alert a human or potentially take steps to secure the network.

Integrated personal assistants

The ultimate commercial vision for a personal assistant is a highly personalised agent that can integrate with multiple systems and help a person manage many aspects of their life. These agentic systems could be embedded in personal devices such as phones. They could also act as a next-generation interface, seamlessly operating across every device and accessed by various methods. We may also see more specialised personal assistants emerge, such as an automated financial assistant that could manage and improve a person's day-to-day finances.

Medical sector

In the longer term, examples might include teams of specialised agents supporting medical diagnosis or helping to create treatment plans. In a care setting, agents may interact with caregivers from many different devices or interfaces to support them with a range of tasks.

Technical developments

Stakeholders emphasised that the pace of change means it is difficult to predict what agentic AI could look like beyond the next two years. Most

stakeholders agreed, however, that we would probably see agents used for a wider range of purposes in a wider range of sectors within this timeframe.

There were conflicting views about how the capabilities of agents could evolve. Generally, stakeholders anticipated that agentic technologies will continue to improve as they are used and tested in the real world. However, they suggested the rate of improvement in LLM capabilities may slow or even stop in the short to medium term.

Multiple stakeholders believed that we could see the following technical developments:

- **Truly multimodal agents:** Many current agentic systems focus on natural language and text-based inputs . For example, a written conversation with a customer service chatbot, or typing a prompt for a research agent. Increasingly, research is focusing on allowing inputs and responses in different ways (eg voice, images and touch).
- **Increasing agent autonomy, and multi-agent systems moving from research to real-world applications**: Most stakeholders highlighted that they expect significant developments in **multi-agent systems** and **agent-to-agent communication in future**. We could see multiple agents deployed to complete separate tasks to meet an overarching goal. Examples might include:
    - research agents in a team working on different parts of a project;
    - agents cooperating with each other to complete a complex task; or
    - agents managing a complex Internet of Things (IoT) system, such as a building.

  We could also see personal agents communicating directly with each other, or with an organisation's agents, on a person's behalf. This would require improved interoperability between agents, which is an ongoing area of research.

- **Agentic AI embedded into a wider range of software and devices**: Stakeholders suggested that in future we could see agentic AI embedded into a wide range of emerging technologies. These might include augmented reality and IoT devices, connected and autonomous vehicles and, even further into the future, robotics.
  Some stakeholders highlighted developments in 'agentic' operating systems. They suggested that future agentic software could be accessible in many different forms and on many types of devices. They also mentioned that people may interact with future agents in their

environment (eg by voice via IoT devices), rather than through a computer or phone screen.
This is consistent with developments we explored in our connected transport and next-generation IoT Tech Horizons report chapters. In July 2025, we published and consulted on our draft IoT guidance.

- **Greater personalisation of agents**: Stakeholders agreed that in the future, agents will increasingly use profiling and have a greater understanding of, and ability to adapt to, a person's environment, preferences and behaviours. An agent could draw on their interpretation of a person's prompts and interactions with the model, as well as information gathered by other connected technologies or applications. For example, a more personalised agent could interpret a person's unique learning style and adapt its content and responses for educational tasks. It could also adapt content in gaming, social media or advertising, or (in the case of a proactive personal assistant) help someone manage aspects of their life.

  This is a development we have explored further in our Personalised AI Tech Horizons report chapter. We have also published profiling guidance.

We could also see the following developments:

- **Emergence of 'self-improving' agents**: Agents capable of rewriting their own prompts to improve performance, or to learn from information they find autonomously as well as the information initially used to train them. Self-improving agents could use this additional information to improve their own models and decision-making. [23]
- **Ongoing research and development in control mechanisms, safety and privacy-focused technical features**: Some stakeholders suggested we could see ongoing research into improvements in control mechanisms, AI safety and privacy-focused measures. For example, this could include research into:
  - real-time ethical autonomy;
  - improving and implementing techniques to mitigate errors; and
  - improving transparency.

  Further strategies to validate outputs may emerge as the autonomy and complexity of agent workflows increase. Some measures may be privacy-focused – for example, future improvements in **privacy-enhancing technologies** (PETs) designed for agents. Others speculated that we may see 'privacy-focused' on-device agents emerge.

In parallel to these developments are approaches that attempt to address some of the limitations of LLMs. [24] Some developers argue that these limitations restrict the development of truly agentic systems. [25] At a high level, agents may appear proactive, able to anticipate and able to learn. However, they remain embedded in approaches that can struggle to turn text-based input and output into action in the physical world.

- **Large action models** (LAMs) offer the means to potentially deliver multi-modal outputs beyond audio, visual and text-based prompts. For example, this may mean the movement or action of a device or robot. LAMs may achieve this through systems such as:
    - action tokenisation (breaking physical actions into discrete sections); and
    - autoregressive action generation (where previous actions are used to predict future actions).

  Much like LLMs, some see LAMs as benefiting from training on very large data sets. Here, LAMs raise similar data protection and privacy concerns as LLM-based agents.
- **Small language/domain-specific models** (SLMs and DSMs) are approaches that seek to address both the size of data sets and the risk of hallucinations within highly technical and specialised sectors (eg law and medicine). [26] These models use smaller, more specialist training sets that reflect technical language in areas with high demands for accuracy and precision of outputs. Their size may also allow SLMs to run on local devices, improving privacy-focused approaches.

While these may come to play an increased role in certain sectors, smaller models are still large compared to early-stage general models such as GPT-5. Furthermore, complex goals may require multiple agentic systems acting together, rather than using a single system. This raises its own concerns. As a result, issues created by **probabilistic** approaches to accuracy remain important, as we explore later in the report.

To deliver improved services, all fundamental processing that underpins agentic approaches (whether LLMs, SLMs or LAMs) must improve its interpretation of, and reaction to, context. Traditional methods of providing relatively scripted responses limit the agent's ability to interact creatively with ambiguous and uncertain situations. Research into **neuro-symbolic AI** seeks to address this.

Neuro-symbolic AI is defined as a broad subsection of AI that seeks to create systems that learn to interact with the world around them without the need

for strict rules or interpretation. [27] Some researchers are developing models to become more 'human-like' in their processes. They do this by combining:

- neuro-connectionist approaches (which seek to mimic mental processes in humans); and
- a focus on symbolic interpretation (assigning meaning to objects, processes and actions). [28]

This approach emphasises the development of self-awareness in systems to allow for better planning, improved adaptability and greater transparency. This would theoretically develop agents that could respond in a far faster and more naturalistic manner to unplanned events occurring around them. There has been a significant growth of interest in this area of research in recent years, but the technology has not yet been used in commercial systems.

While neuro-symbolic approaches offer gains in accuracy and independence, they are also likely to raise heightened concerns around:

- a growing demand for contextual and environmental information to inform the development and decision-making of neuro-symbolic processes; and
- a lack of transparency about why agents that interpret situations creatively make the decisions they do.

Both issues are examined in greater depth later in this report, in the data protection issues section.

We do not attempt to provide a judgement about whether LLMs and the systems they support can reason in the creation of their outputs, or if they produce outputs that predict what reasoning might look like. This report operates on the premise that agentic AI outputs can be presented as reasoned and may often be interpreted and used as such by organisations and people. Whether or not this use is appropriate, it raises considerations for privacy and data protection.

1 OpenAI article on introducing Operator agent; Anthropic article on developing a computer use model; Google article on Gemini AI

2 Financial Times article on Open AI hopes for AI agents; MIT Technology Review on helpful AI Agents; Axios article on uses of generative AI

3 Anthropic news release announcing Claude Opus 4.1 (August 2025); Anthropic news release introducing Claude Sonnet 4.5 (September 2025)

4 Google news release introducing the Gemini 2.5 Computer Use model (October 2025)

5 Anthropic news release on building agents with the Claude Agent SDK (September 2025); OpenAI news release introducing AgentKit (October 2025)

6 Microsoft Support article on experimental features; Windows Central article 'Microsoft just revealed how Windows 11 is evolving into an agentic OS' (November 2025)

7 Salesforce news release announcing "the Agentic Enterprise" (October 2025)

8 GOV.UK Artificial Intelligence sector study 2023

9 GOV.UK article on potential uses for AI Agents

10 There is an ongoing debate about the extent to which this is 'reasoning', and we do not adopt a position on this debate.

11 Arxiv article on characterizing AI Agents for alignment and governance; Arxiv article on why fully autonomous AI Agents should not be developed

12 Arxiv article on how underspecification presents challenges for credibility in modern machine learning

13 OpenAI article on "Buy it in chatGPT"; Forbes article on AI shopping agents

14 Meta AI tools for businesses marketing on social media; Salesforce article on agentic commerce

15 PwC article on the rise of agentic commerce on buying behaviour; Salesforce article on agentic commerce

16 IBM article on agentic workflows

17 Salesforce article on Agentic AI; Recruiter article on AI Agents

18 Simplifai article on AI and insurance

19 See, for example, Personal AI portfolio of AI personas

20 GOV.UK article on potential uses for AI Agents

21 Opening Remarks by Minister Josephine Teo (Government of Singapore) at Google Cloud's AI Asia Event

22 There are early examples of this. For example, in November 2025, Anthropic released a statement about their discovery of a threat actor using agentic capabilities to execute a cyber attack: Anthropic article 'Disrupting the first reported AI-orchestrated cyber espionage campaign'

23 Arxiv article on Self-Adapting LLMs; Google DeepMind article on Alpha Evolve

24 IBM article on Agentic AI

25 Arxiv article on Large Action Models

26 IBM article on Small Language Models

27 Arxiv article on Neurosymbolic AI

28 University of Edinburgh article on Neurosymbolic AI

Data protection and privacy risks

Many of the data protection issues with agentic AI applications are similar to those raised by other types of AI, and in particular, generative AI. In some cases, however, the characteristics and capabilities of agentic AI could exacerbate existing data protection issues or introduce new ones. We are starting to see these risks emerge with current agentic AI systems. We anticipate these risks growing with increasing capability and adoption of agentic AI, unless efforts are made to mitigate them.

Data protection is part of a wider set of considerations when new and emerging technologies are being developed, and how data protection law is applied will affect social and economic outcomes. Whilst our focus is data protection issues, we know that these are not isolated from society and the economy. Potential harms and risks from agentic AI such as the following are out of scope for this report:

- the economic impact of job losses;
- environmental impacts;
- impacts on competition; and
- collective disempowerment.

However, we will continue to ensure we are working with stakeholders to understand where data protection touches upon these issues.

Human responsibility and controllership

Despite the language of ‘agency’ and ‘agents’, and the hype around agentic AI, agentic AI systems will not be conscious or have intent within the next two to five years. They are not and should not be considered as legal entities, even if organisations using agentic AI may seek to blame it for errors. In the context of data protection, AI agency does not mean the removal of human, and therefore organisational, responsibility for data processing. Organisations must be clear on the expectations that still apply under data protection legislation. They remain responsible for using personal information in an appropriate fashion.

Today, many agents under development give the organisation or person deploying them control over factors such as:

- **the actions the agent is authorised to take**. For example, a person could authorise an agent to automatically make purchases up to a certain financial amount. The agent would need to seek permission to make purchases above that amount. A person could also configure an agent to automatically block illegal purchases; and
- **the information the agent can access**. For example, an organisation could configure an agent to only access specific internal information for certain tasks, rather than calling on external tools. Or they could configure an agent to request permission before accessing certain personal information.

Many current agents also have features that allow humans to review what they are doing in real time. Examples include the ability to watch an agent navigate a browser, or a window where the agent describes what it is doing.

However, the future extent of human involvement in agentic AI actions is unclear. As the level of agency, the unpredictability of the environment and the software's ability to learn increase, it becomes harder to fully specify what a user wants an agent to do. Using LLMs to manage context and assumptions and allow natural language interfaces is part of this advance, taking agentic AI beyond simple automation. LLMs allow for tasks to be described much less specifically than traditional programming approaches, which require programmers to set out steps in detail.

This new development means that an agentic system may misinterpret instructions or do something unexpected. In multi-agent environments, agents may be communicating, collaborating and possibly even colluding. [29]

At best, failures result in poor performance of the agentic system, and at worst, the user of the system or third parties may experience loss or harm. This is likely to be exacerbated by the fact that, by design, agentic systems operate with limited human involvement. The intentional or gradual reduction of oversight may mean that these unintended consequences and harms persist for a longer period before being noticed and rectified. To mitigate and minimise these instances, governance systems have been developed and proposed, including measures to preserve privacy.

Increasing agency means that developers and deployers of agentic systems don't have full control over the behaviour of those systems. But they retain control over whether to deploy those systems or not, and the level of risk they tolerate in doing so.

Some experts argue that we should not develop fully autonomous AI – systems capable of writing and executing their own code beyond predefined constraints – because doing so significantly increases risk. [30] Others say that full autonomy for AI systems is not a desirable goal, precisely because it means a loss of control for the deployer or user. These commentators note that this might be avoidable if the proper controls and constraints are put in place. [31] It will be essential to identify one or more data controllers to ensure these controls are effective.

Governance

Measures to govern agents and agentic systems must be flexible enough to handle changes in priorities, goals, tasks and the environment in which the agent is operating. They must also consider how these systems might develop in future, as capabilities and functionality advance. Proposed governance schemes include the Safer Agentic AI Foundations, from the Agentic AI Safety Community of Practice. This scheme covers an end-to-end overview of securing agentic operations, with a heavy focus on continuous documentation, monitoring and review.

Furthermore, organisations are likely to retain many of these governance responsibilities. Placing the sole responsibility for creating, applying and maintaining these value-derived governance frameworks on the end user would not be universally applicable or suitable. In the case of agents available to the public, it is unlikely that members of the public have the knowledge or skills to:

- apply or update these frameworks without help; or
- address any issues, particularly if an agentic system is working on tasks or at speeds or complexities which may be hard to understand.

This highlights the responsibility of suppliers of those systems, both to:

- employ good governance before the point of sale; and
- ensure that the systems they provide are suitable for customers and their tasks.

Automated decision-making

Developers and organisations may include automated decision-making (ADM) in agentic AI systems as they seek rapid automation of increasingly complex tasks. Effective ADM implies minimal human involvement in the decision.

However, that may have significant or legal effects on them. This is part of the ADM provisions.

The legislation also requires people to be informed about important automated decisions made about them. It gives them the right to contest those decisions and ask for a human to intervene in the decision-making.

Organisations using agentic AI to make decisions about people would need to consider:

- how the decisions impact people;
- how to clearly communicate the use of automation to the people affected;
- how to put in place systems that allow people to contest decisions; and
- how to effectively and meaningfully allow humans to intervene in agentic AI decision-making.

Organisations building or deploying agentic AI must be aware of their obligations under data protection legislation.

We will provide clarity on our regulatory expectations around agentic AI as part of the development of the code of practice on AI and ADM. We are planning to release an interim update on ADM and Profiling at the beginning of 2026.

Purpose limitation and data minimisation

Purpose limitation

- they are clear and open about why they are collecting personal information; and
- what they intend to do with the personal information is in line with people's reasonable expectations.

This is important because there is a risk that organisations could set purposes too broadly in order to encompass all potential operations.

Organisations must:

- have a clear purpose for collecting and processing information used by an agentic AI system; and
- communicate that purpose clearly.

This includes information collected during the operation of the system as well as during the development phase.

As with other types of AI development, agentic AI development and use involve different processing stages. Each stage involves processing different types of personal information for different purposes. Having a specified purpose in each stage allows an organisation to:

- understand the scope of each processing activity;
- evaluate its compliance with data protection; and
- help them evidence that.

We have published a response to the call for views on generative AI and how the purpose limitation principle should be interpreted within its lifecycle.

Organisations should take appropriate technical and governance measures to ensure agentic AI systems operate appropriately, using information in a way that people can reasonably expect. We explore this further below. Organisations looking to deploy agentic AI and systems should be aware of the guidance that we are producing in this area.

Data minimisation

Under article 5(1)(c) of the UK GDPR, personal information must be:

> adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimisation).

While we provide guidance on data minimisation, our stakeholder engagement and research highlighted diverging perspectives in this area. Many emphasised that granular user controls are important. This refers to an organisation or person choosing upfront what information and systems an agent is permitted access to for a specific purpose.

Some stakeholders suggested that people may progressively choose to open access to additional information, once the system has proven it functions correctly. Others favour a more open approach while acknowledging the importance of controls, particularly in specific contexts where agents have access to a wider variety of information upfront. This approach considers that

access to context helps an agentic system learn and perform better and deliver more personalised results.

Stakeholders noted that they expect these issues to compound as the technology evolves. Developers may design some agentic AI systems to be open-ended (rather than highly specialised) and flexible in approaching problems. With growing demands to personalise systems, this may create challenges for complying with data minimisation.

Data minimisation is closely related to the purpose limitation principle explored above. Once an organisation using an agentic AI system defines the purpose for using it, it can establish the type and volume of information needed to fulfil that purpose. Data protection law requires organisations to process only the personal information needed to achieve the specified purpose.

Organisations should not give agentic AI systems access to information just because it might be useful in the future. They must have a justifiable reason for the agentic AI to access and use the information.

It is currently unclear what technical measures may be implemented in future to achieve both effective delivery of goals and appropriate processing. However, our stakeholder engagement highlighted that organisations are already considering how they approach data minimisation as agents evolve.

Careful selection of the tools and databases an agentic system has access to will be important. Developers can also design systems with other controls, such as asking a human for permission when an agent needs to access personal information. Other approaches (eg masking of personal information, age verification, system permissions, observability techniques and transparency notices) may support this and are expected as part of good governance practice.

Some stakeholders have noted that lack of access to enough data, systems and tools would significantly limit what agentic systems can do. However, compliant future development of agentic systems will rely on organisations developing and implementing tools to effectively manage access to personal information. This process is equivalent to the principle of least privilege used in many organisations.

Rapid generation of personal information by agentic AI systems

As well as processing large quantities of personal information, agentic systems may create additional data protection issues through their ability to rapidly infer and create new personal information at scale. In some instances, this may constitute profiling if it evaluates certain personal aspects relating to a person. Furthermore, when agentic AI systems make legal or significant decisions, those decisions may be considered to be ADM. This raises issues explored above and discussed in our recent call for evidence about generative AI.

Organisations will need to effectively and appropriately manage significant and growing quantities of personal information to ensure that they are compliant with data minimisation requirements. This involves:

- meeting expectations of effective technical and organisational measures;
- upholding data rights; and
- complying with data protection requirements (eg storage limitation, security and data minimisation).

Agentic AI systems increase the chance of 'cascading hallucinations'. Hallucinations occur when an AI agent generates inaccurate but plausible information (potentially including personal information). The agent may share this information through:

- the tools or databases it has access to; or
- interactions with other agents.

This may cascade the inaccurate information across multiple locations or through the stages of a decision-making process. In such circumstances, compliance with the accuracy principle of data protection may become harder. It could also significantly raise the potential for serious harms to occur when processing sensitive personal information or services.

Special category data and agentic AI

Agentic systems may draw upon or generate sensitive information, such as special category data , in unexpected ways in the pursuit of open-ended goals. While some deployments of agentic AI may use special category data by design (eg the management of healthcare records and treatment decisions), other cases may not be obvious from an initial goal. The organisation is responsible for considering:

- the purpose of processing;
- the users' preferences (where the organisation can tailor preferences to deliver different levels of service); and
- when this may include the processing of special category data , which involves enhanced requirements under data protection legislation.

It is possible that, even with a purpose that does not involve processing special category data, agentic systems may still use information to infer and use special category data to make a decision. Organisations should assess whether their agentic system has the potential to infer and use special category data in pursuit of its purpose. If it does, organisations should ensure:

- that they have a valid lawful basis and an article 9 (of the UK GDPR) condition for this processing; and
- that people are aware of the possibility of inference and use of their special category data.

Alternatively, organisations could consider technical measures to restrict the agentic system's ability to infer and use this type of data.

Explicit consent as a basis for processing special category data may prove difficult to achieve when deploying agentic AI and systems unless people have a genuine choice – for example, in situations in which the agentic system's services can be used appropriately without special category data or with greater personalisation if the user chooses.

This is already an established issue for wider forms of AI such as LLMs. The highly complex data flows of agentic systems are likely to make it harder to meet the expectations necessary for transparency. They also complicate the enactment of rights, such as removing information when a person withdraws the consent that is necessary for explicit consent.

Organisations should review our guidance on special category data andconsent as a basis for processing.

Transparency and explainability

The transparency principle requires organisations that process personal information to be clear, open and honest from the start about:

- who they are;
- how and why they use people's personal information; and

- the data rights available to these people.

See our guidance on lawfulness, fairness and transparency.

The principle applies even if the organisation has no direct relationship with the people whose information they are collecting, or when they collect the personal information from another source. If organisations are not transparent, there is a risk of people being unaware that their information is being processed. This is called 'invisible processing' and could prevent people from being able to exercise their information rights.

Complex information flows make the potential development of agentic AI particularly relevant as these systems grow in capability. In this ecosystem, information may be:

- shared at different stages, among different agents; and
- developed by different organisations, to make different decisions.

As with non-agentic AI (and all processing activities), any lack of transparency and explainability in foundational models or agent actions may result in unintended harms to the people affected.

As the capability of future agentic systems increases, there is a risk that these systems may autonomously generate new ways of using personal information. For example, they may:

- autonomously process information they need to perform a task, which might be beyond what developers anticipated;
- process personal information to pursue its human-set objective in a way that the person the information is about may not reasonably expect;
- seek out new personal information without a person's knowledge; or
- re-purpose information originally collected for a different purpose.

There are already cases of AI agents acting in unexpected ways. [32] These issues could result in a system making decisions in ways an organisation did not foresee at the time of the system's deployment.

The growing emergence of agent-to-agent communication that isn't visible to human observers may also significantly impact transparency. This is because it will become increasingly difficult to understand how and where information is processed. It could also impact the appropriate implementation of human intervention and data rights (both explored below).

While the autonomous nature of future agentic systems could pose challenges to organisations for transparency, the organisations' obligations as data controllers remain the same. To ensure that personal information is processed transparently in agentic systems, organisations must consider their obligations and how they will meet them before they begin processing.

In line with our guidance on data protection impact assessments (DPIAs), if organisations assess that deploying agentic systems will result in a high risk to people's information, they must carry out a DPIA. Organisations must provide relevant privacy information when they collect personal information directly from people. If they collect the information from other sources, organisations must provide privacy information to people within a reasonable period. See our guidance on ensuring transparency in AI.

Accountability

Linked to the need for transparency is the organisation's ability to trace the use of personal information and provide accountability. Technical approaches already exist that enable organisations to review what an agent is doing either in real time or retrospectively. Some stakeholders emphasised that existing explainability approaches are still relevant to agentic systems. Accountability-based approaches are expected to evolve. However, it's not clear whether, in practice, organisations will be able to effectively monitor an increasing number of agents. They may, for example, need to rely on supervisor agents.

Organisations must give further consideration to the dual demands of performance and accountability. For example, one potential way to increase an AI agent's statistical accuracy is to create multiple, focused agents trained for specific tasks. However, this increased technical complexity (eg multiple opaque tool calls or unplanned execution) could make it much harder to understand:

- what data the system is using; and
- how and where it's making decisions.

This might reduce transparency for people about how, where and why organisations are using their information, which would limit their ability to exercise their data rights.

Accuracy

Article 5(1)(d) of the UK GDPR states that personal information must be 'accurate' and rectified promptly where this is not the case. While the UK GDPR does not define 'accuracy', the Data Protection Act 2018 says that 'inaccurate' means "incorrect or misleading as to any matter of fact".

Approaches such as **chain of thought** and **retrieval augmented generation** (RAG), as well as further context-specific fine-tuning, can enhance the accuracy of LLMs. However, the fundamental issue remains that these models predict text in a probabilistic manner – they don't logically deduce findings from the world and through reason. This leads to the widely explored issue of hallucinations. A hallucination is where LLMs describe (often convincingly) invented or misinterpreted facts.

As noted above, increased statistical accuracy is a key focus of emerging models and systems, although how this will be achieved is unclear. Even if this can be realised, the more immediate issue remains that different contexts may require different levels of statistical accuracy depending on the risk of harms. How these situations can be monitored and responded to appropriately is a key issue for organisations using agentic systems.

The matter of accuracy may become increasingly important in agentic AIs as they come to make greater use of previous information, and decisions made using personal information, to achieve a goal. Stakeholders have noted that inaccurate information held in an agentic system's memory (either acquired or hallucinated) is likely to have a significant effect across multiple decisions. Whether agentic AI systems hold inaccurate data in long- or short-term memory can also define how complex the technical solution may be. Solutions can range from a simple reset of preferences to more fundamental fixes in a model. This may lead to risk of harms to users, from simple but easily correctable errors to significant decisions being incorrectly made.

We have already provided regulatory certainty about problems that these hallucinations can cause from a data protection and privacy perspective. See our response to our call for views relating to generative AI.

At a high level, hallucinations risk large-scale generation of inaccurate information that can become rapidly embedded in systems. These can lead to opaque and unfair decisions that might cause harms, including financial loss or even physical or legal harms. Agentic systems risk supercharging this issue. Given the increased number of actions and decisions they require to achieve a goal, they could reduce the opportunity for meaningful human intervention (and correction).

Individual information rights and fairness

Organisations developing and deploying agentic systems must be aware of the requirement to implement data protection by design if those systems will be processing personal information. Under UK GDPR, people have various rights, including:

- the right of access;
- the right to erasure;
- the right to rectification;
- the right to object;
- the right to portability; and
- rights about automated decisions.

Organisations processing personal information in agentic systems must consider how this processing could impact people's individual rights. Organisations should implement data protection by design. Otherwise, the opaque data flows, ADM and multi-agent interactions that characterise advanced agentic systems may make it harder for people to exercise their rights.

For example, without fit-for-purpose technical and organisational measures, the complex information flows within a future agentic system could make identifying information held for an individual user increasingly difficult. There is a risk that this could make it more difficult for people to exercise their right to access copies of their information.

Unless organisations take appropriate measures, the growing complexity of agentic systems may also impact people's ability to exercise their right to rectification (correction of inaccurate personal information). The complexity could make it harder to determine the source of personal information. For example, an agentic system used by a lettings company might assess the reliability of a potential tenant based on incorrect information. An existing system may draw information from data sources approved by the organisation. However, a future agentic system could autonomously draw information from a vast number of data sets. It could also use these sources to create inferences about the person. This means the original error and its effects may be difficult to both identify and amend.

Organisations deploying agentic AI must uphold individual rights and design agentic systems with data protection by design and default. This will enable transparency, accountability and meaningful human intervention.

Fairness

The principle of fairness in data protection is based on the premise that organisations should not process personal information in a way that is unduly detrimental, unexpected or misleading to the people concerned. The ADM processes that agentic systems use may diverge from the original scope the organisation envisioned as a system reacts to and learns from its environment. Organisations must take care that agentic systems do not process personal information in ways that people do not expect in practice. See our guidance on lawfulness, fairness and transparency.

The role of the data protection officer

The data protection officer (DPO) monitors an organisation's compliance with relevant data protection laws. They also advise on data protection obligations and provide insight on DPIAs.

Many of the data protection obligations about the use of agentic systems are the same as with other AI applications, and we have previously published guidance on AI and data protection. Integrating agentic AI into an organisation has the potential to create novel risks and opportunities for DPOs. However, controller-processor obligations remain applicable to organisations deploying autonomous agentic systems.

Challenges in maintaining oversight over novel processing

It may become difficult for organisations to maintain DPO oversight over new processing if they are encouraging staff to experiment with agents. For example, employees trying out new uses for more autonomous agents in their day-to-day work may quickly end up processing personal information in various unanticipated ways. The use of shadow or short-term AI agents (ie agents spun up for immediate tasks and then torn down just as quickly) further compounds the oversight considerations.

The challenges become more acute if agents:

- have freedom to access a variety of personal information held by the organisation; or
- are permitted to draw on external sources outside of the organisation's controlled data sources and systems.

Multi-agent systems raise a problem of compound loss of privacy. Multi-agent interaction across multiple systems (rather than within them) is likely

to amplify issues of accountability, statistical accuracy and security risks. This may persist even when organisations have taken best steps to ensure data protection by design. Varying approaches to security, transparency and interoperability could lead to multi-agent use significantly increasing the risk of data breaches.

If DPOs and other governance teams develop effective governance structures and well-defined parameters for employees using agents, this may mitigate some risks. The increasing integration of agents across an organisation, the potential expansion of activities they undertake and the significant security risks could require greater collaboration between the DPO and chief information security officer (CISO).

Increased complexity of documenting decision-making

Agentic systems may be operating at pace, and they can be susceptible to goal or decision-making 'drift'. As an agent responds to its environment and learns from the information it collects from the data sources it accesses, it may need to document its decisions as part of governance processes. This documentation would need to be readable, verifiable and accurate.

Organisations may need a separate, standalone monitoring system (or agent) to monitor logs, interpret them and intervene as necessary. This could enable the organisation to verify that the system is still acting according to its governance structure. The autonomous nature of agentic systems adds further complexity to this requirement, because systems will be making decisions without being directly observed.

Evolving role of the DPO

Some researchers have predicted that the control and supervision of AI agents could increasingly become a part of people's jobs. If an organisation widely uses agentic AI, it is plausible that the role of the DPO could also evolve.

We already see a degree of automation for tasks (eg subject access requests, cookie consent management or breach reporting). There is a significant existing market for privacy technology. This is increasing in regulatory technology sectors, such as financial services. Agentic AI could extend this further, through the addition of so-called 'virtual employees' supporting the human DPO.

One possible future envisions 'DPO agents'; systems integrated into privacy teams to scale and augment the role of human staff. Organisations could use them to:

- upscale oversight of the data and processing in an organisation;
- scan proactively and flag up new or high-risk processing activities across an organisation;
- help with specific DPO tasks (eg reviewing, identifying and assessing high-risk suppliers based on internal governance criteria); or
- detect breaches by analysing relevant regulations (domestic and international), identify compromised data, and suggest next steps for and collaborate with DPOs.[33]

In such an arrangement, the future role of a human DPO could shift towards orchestrating and managing a team of 'data protection agents'. This could include defining the boundaries of what an agent can or can't do and what systems it has access to. This future envisages DPOs drawing on agent insights to focus their attention and engagement on an organisation's highest-risk processing.

Agentic AI security threats and mitigations

Data protection law requires organisations to protect the information they are processing by means of appropriate technical and organisational measures. See our guidance on data protection principles. The security principle of data protection applies to any personal information processed by agentic systems. The security principle also requires organisations to ensure the "confidentiality, integrity and availability" of information processed.

Agentic AI, as with any other system connected to the internet, may be subject to attack by malicious third parties. The autonomy of agentic systems and their ability to perceive and learn from their environment presents novel opportunities for compromise. This puts at risk any personal information held within and processed by these systems.

The features of agentic AI and agentic systems that differentiate them from other forms of AI present potential novel opportunities for attack, for example:

- distorting goals;
- attacking the reasoning of an agent;
- manipulating reasoning; or

- poisoning data in the system's memory.

Potential taxonomies of threats and proactive and reactive security measures are already under development, including the Open Web Application Security Project's threats and mitigations list. These identify new attack surfaces that agentic AI might introduce into an organisation, as well as documenting security from first principles to the eventual shutdown of the system.

Agent business models and the concentration of personal information

Some potential privacy risks from agentic AI arise from business models and product design decisions rather than the inherent nature of the technology. Personal assistant agents are one possible model for products based on agentic AI. These agents are frequently envisioned as general-purpose assistants that handle a wide range of personal tasks and act as an interface between the user and the digital world.

To be effective, this type of agent might need access to information about a person, their preferences, their environment and previous behaviour. This may also include information about third parties the user interacts with. It would also need access to digital tools (eg communications, calendars, accounts, IoT devices, and possibly even the user's ID). This might involve personalisation, combining information from many separate services and granting the agent access to secure services.

However, this has the potential to undermine data protection controls, including security and encryption. This creates an opportunity for the agentic system to accumulate extensive personal information, increasing the risk of surveillance and data breaches. As we do not anticipate most people would create their own agents, they would need to source agentic assistants from third-party providers.

Demand for personalisation may lead to the embedding of personal information within models. Should that happen, and the models are either publicly available or become shared more widely, there is a risk that third parties may extract the embedded personal information from that model.

29 Arxiv article on secret collusion among AI Agents

30 Arxiv article on why fully autonomous AI Agents should not be developed, p.1

31 Arxiv article on characterizing AI Agents for alignment and governance, p.8

32 BBC News article on AI Agents going rogue

33 See, for example: An announcement from Onetrust about its first data privacy breach response agent

Innovation opportunities – What innovation might the ICO want to see in agentic AI?

We would like to identify, encourage and support opportunities for innovation within agentic AI that support data protection and information rights.

Taking advantage of these opportunities may require related developments in the capabilities and reliability of agentic AI, as well as appropriate testing and risk assessment. We see this as part of moving us closer to the high-capability scenarios earlier in the report. Each of these contexts requires thinking about how to develop innovations safely and responsibly.

We will address innovation opportunities proactively as agentic AI matures and our role in regulating it develops. As part of this, we will:

- invite developers to use our Innovation Advice service. This is a free, fast and direct service for organisations doing new or innovative things with personal information. It can give advice to help solve the data protection issues holding up the progress of new products, services or business models; and
- continue to invite innovators to work with our Regulatory Sandbox to engineer data protection into these technologies from the outset, focusing on the most innovative propositions.

Additionally, we have identified several areas where innovation in agentic AI could contribute towards privacy-positive outcomes. These are discussed below.

Data protection compliant agents

Agent governance systems typically focus on aligning agent norms with 'human values'. Some of these systems focus on legal, moral or ethical bases. There are fundamental questions in responsible AI about:

- which values get built into AI systems;
- how these should change over time or across jurisdictions; and
- how they are assessed.

Agentic systems must never take actions that break the law. Data protection compliance is a legal requirement and a minimum threshold that all developers and deployers must comply with.

Part of the novelty of agentic systems is the use of agentic AI for planning tasks. As well as agentic systems being legally compliant, we are interested in approaches to:

- ensuring that they generate outputs (eg plans or proposed solutions to tasks) that comply with data protection law; and
- how to embed legal obligations in agentic AI (eg data protection by design and by default) at a fundamental level.

Such approaches to responsible data protection compliance could be a market differentiator for those AI agents and agentic systems that can demonstrate it.

Agentic controls

The governance of agentic systems will require a range of tools to manage intellectual property, cybersecurity or public relations. Such tools will likely also play a role in supporting data protection. These tools could include tools for monitoring, auditing, explainability, permission structures, authentication and data access protocols. Valuable areas of research would include:

- the extent and types of personal information storage and processing needed to deploy and govern agentic systems;
- methods for privacy-enhancing agentic governance;
- privacy-enhancing communication between AI agents; and
- mechanisms for redress and restitution, particularly in agent ecosystems – ways to understand how something has gone wrong and how to put it right.

Layered or tiered protections, which cascade from fundamental protections implemented by suppliers, could incorporate principles such as privacy by design and default, or security-first design. Organisations could build organisation-specific controls on top of these by setting out which information sources the agents could access, and with what permissions. Protections implemented by individual users could give them granular control over what information is shared with the agentic system.

Privacy and personal information management agents

Some people find actively managing their personal information and protecting their privacy difficult. It can be complicated, made harder by dark patterns, and can go wrong in ways that are not easily apparent to the person involved. Our work on damaging website design practices and research on cookiesshow that this often leads to outcomes that diverge from people's desired privacy choices.

We are keen to explore the possibility of developing AI agents that take on some of this burden, for example, by interpreting website privacy policies or cookie statements and comparing them to their users' preferences. In survey research we commissioned, two in five UK adults admitted to never reading cookie policies or settings. In this context, an agent designed to be a more vigilant guardian of personal information than a person themselves would be an impactful innovation.

Local agents and trusted computing

Agentic AI may be part of the solution to keep transactions and tasks confidential. For example, a correctly configured agent might be able to act as a proxy for a real person who wants to remain confidential.

We anticipate innovation opportunities for agents to conduct research tasks where information is held locally (eg on a device) and report back their findings (perhaps a yes or no decision for an application) to their user. For example, a person's agent may be able to process several contract offers and pick an appropriate one, without providing personal information to the potential providers. These applications would require the combination of agentic approaches with privacy-enhancing technologies.

Freedom of information and data protection agents

Compiling responses to freedom of information (FOI) requests can be time-consuming and complex for organisations. As agents develop, they may become more skilled at processing large amounts of information in public records. This could allow organisations to identify relevant information more quickly. Agents could also:

- help in triaging by categorising incoming requests and flagging them to relevant parts of the organisation; and
- monitor organisation response times and help in complying with deadlines.

Agents could play a similar role in responding to requests related to data protection rights. They could help organisations search for and identify relevant information. This would help organisations in responding to requests from people seeking copies of their information or requests for correction and deletion. Developers of such technologies would need to consider how to mitigate hallucinations. Deploying organisations would need processes to ensure that they did not provide people with hallucinated or incorrect information.

Benchmarks and evaluations for agents

Benchmarks are standardised tests of evaluation criteria used to measure or compare the performance of an AI system across specific tasks or capabilities. AI developers use benchmarks during model development to refine and compare their models. There are competitions to score highly on key public benchmarks.

Third parties can use benchmarks when selecting between different AI tools. Benchmarks can measure an AI system's ability to solve math problems, orchestrate complex tasks or recognise images. For AI agents, benchmarks involve more complex tasks that assess reasoning, the ability to handle different forms of information and the ability to use tools. Designers often design AI agents to simulate real-world tasks.

We would welcome innovation in methods for the practical evaluation of the compliance of agentic AI systems with data protection legislation. This would ideally include both:

- compliance of the agentic system itself (eg how the organisation set it up, what training information it uses or what personal information it processes); and
- compatibility of any actions that the system advises or takes with data protection law.

More broadly, research on which tasks agents can and cannot perform well would enable organisations to make informed decisions about choices and risk assessments when deploying agents.

Scenarios for the future of agentic AI

Scenario planning

In this section, we use scenario planning to explore the potential near future of agentic AI development.

Scenario planning is a foresight method that aims to identify and manage key uncertainties. This involves creating a range of plausible future scenarios that encompass that uncertainty but still support flexible planning. Between them, the scenarios provide indicators about the direction of the near-future development of the technology.

We began by identifying drivers and trends to create scenarios which would describe possible futures. These allow us to examine some of the opportunities and risks agentic AI might present to data protection in those futures. These drivers and trends impact how agentic AI might develop, and they include technical considerations such as:

- the cost of training AI models;
- a drop in the cost of the compute power needed to run agentic systems (because of advances in hardware and improvements in AI algorithms); and
- the availability of datasets to train AI.

We also considered social and economic factors, including:

- the amount of investment into agentic AI;
- an organisational fear of missing out or being left behind (driven by the hype cycle and marketing); and
- the draw of potential organisational savings from reduced staff and labour costs.

During our research into trends driving adoption, we also saw indications that the adoption of agentic AI could follow a similar path to the proliferation of previous AI technologies (such as generative AI and LLMs). This could include its integration into other platforms (both online and on-device). You can find more detail on some of these trends in Annex II.

We selected two variables to create a four-by-four scenario grid. We chose the capabilities and the extent of adoption of agentic AI. Good variables for

scenario planning should be uncertain (both high and low ends should be plausible). They should also have a high impact in terms of agentic AI, its information rights impact and its regulation.

The capabilities of AI agents. Researchers have identified a wide range of characteristics for evaluating agents and other tools. Capability is a general term we use to reflect a range of component characteristics, such as:

- autonomy;
- generalisability (ability to work across a range of different tasks);
- ability to handle underspecified tasks;
- goal complexity;
- alignment;
- reliability;
- efficacy; and
- the extent to which researchers have addressed the current limitations of LLMs.

We acknowledge that various capabilities may not increase at the same time. For example, we might see great progress in the speed of agentic AI, but no great advances in its reliability. For our purposes, a variable that bundles together capabilities is sufficient.

Capability is an important variable because of uncertainty and the significant hype around agentic AI. Many stakeholders told us that agentic AI is inevitable, but also discussed significant technological barriers to implementation. Some mentioned that the controls in place and choices made by organisations or people implementing them can impact a future agentic system's capabilities.

There are already diverging approaches about what an effective, capable agentic system might look like in practice, and whether this is built around specialised, tightly controlled agents or general-purpose agents. Capability is likely to make a large difference to the social and economic impact of agentic AI and how it should be regulated. While the low end of capability is close to the status quo, the high end represents a step-change in technology.

At the lower end of this scale, we position AI agents that are not much more sophisticated than current chatbots and ADM tools. They likely remain based on LLMs and share most of the limitations of those models. The ability to learn and interact effectively with the environment may also remain limited,

minimising physical and multi-modal outputs. AI agents remain task-specific and prone to unpredictable failures.

At the higher end of this scale, we anticipate technological developments to expand the capabilities of AI agents in many areas. They are capable of handling more complex problems, acting with more autonomy and in a wider range of contexts. Even at the higher end, we do not assume that agentic AI will unlock radical advances in AI research and development, causing an exponential increase in capability.

High capability is not superintelligence. However, in high capability scenarios, some agentic systems will be able to write and edit their own code. There will be tasks to which agentic AI is better or worse suited, and many areas where using agentic AI is inappropriate.

With capability, we are talking about the capability of agentic systems as a whole, not just the raw potential or abilities of agentic AI models. As set out in our scenarios, we assume that in the high capability scenario, some advances have been made in managing and controlling agentic AI systems. This reflects existing efforts and the practical desire from organisations to have reliable technologies.

Adoption of agentic AI. This variable captures the extent to which society will use agentic AI.

A low adoption scenario would see agentic AI used in limited ways, perhaps in certain sectors or for specific tasks, but not commonplace. A high adoption scenario would see agentic AI used nearly everywhere, and agentic capability integrated into many other systems. Our scenarios assume that AI will spread unevenly across all use cases and sectors.

The current hype around agentic AI suggests that adoption is inevitable; however, not all new technologies achieve high adoption. There are various reasons for this (eg vulnerabilities, lack of demand or inability to find a successful business model). Technical adoption can also be slow, especially when meaningful adoption requires changes to business and organisations, or even to social norms and laws. At present, we have seen few examples of large-scale deployment of agents, but this could change. Even for generative AI without agentic scaffolding, adoption in safety-critical areas is low, with simpler models preferred.

Adoption is also significant for how we regulate. A low adoption scenario may mean that we rarely have to consider agentic AI. A high adoption scenario means it will frequently be a factor in most of our work. Rates of adoption of agentic AI and its concentration across sectors will therefore be an important metric for us to follow.

This combination of two variables gives the following **high-level scenarios** for the near- to mid-term future of agentic AI:

	Low agentic capability	**High agentic capability**
High adoption	Just good enough to be everywhere (scenario two)	Ubiquitous agents (scenario four)
Low adoption	Scarce, simple agents (scenario one)	Agents in waiting (scenario three)

Table 1 (above): Four scenarios for the future of agentic AI

These four scenarios formed the basis of our analysis of privacy and data protection considerations as they might present in future developments. The issues identified within those scenarios will help inform future policy thinking. They are presented here as a framework for us and various stakeholders to consider the safe development of innovative agentic AI applications.

The following scenarios aim to explore possible developments and uses of personal information by agentic AI. While the scenarios include high-level commentary on aspects of relevant data protection compliance, you should not interpret this as confirmation that the relevant processing is either desirable or legally compliant. This report does **not** provide ICO guidance.

Scenario one: Scarce, simple agents (low adoption, low agentic capability)

In this scenario, we see low adoption of agentic AI because of limited capacity. Agentic capabilities remain close to the current state, and the challenges of LLMs remain. Agentic AI has not lived up to the marketing hype about its potential. Software providers may offer agentic AI, but people and organisations don’t systematically use it. We would expect to see the following:

- Growing suspicion of agentic AI’s accuracy and reliability means it is not widely used commercially and in government. Its use is mostly in low-risk situations where the limitations do not have a serious impact. AI

agents remain as toys, experiments and demonstrations, with use mainly by early-adopter enterprises and research labs. Most organisations do not use agentic AI workflows or restructure their workflows around agents. There is very little use in high-stakes tasks or highly regulated domains.

- Wider society has little exposure to agentic AI. Data protection impacts are not fundamentally different from those of generative AI. However, public awareness of how agentic AI works and where it is used is limited.
- Without widespread use, there are low demands for standardisation, and AI agents are often incompatible. There is little infrastructure set up to support the use of AI agents (eg protocols for agent-to-agent communication, easy agent hosting platforms or data intended to be easily accessed by agents). Similarly, users may deploy experimental agentic AI with immature governance and security practices or little attention to data protection. This reinforces concerns about their use. There is a heavy use of human supervision (eg checking outputs or human approval being required for access to particular tools) as a safeguard against failures of agentic AI.
- Where agents are used, they have limited access to additional tools and databases and limited ability to make decisions on their own. Human supervision, intervention and error correction are required to get practical and useful results from AI agents. Each deployment is relatively isolated, limiting the risks of large-scale failures or cascading problems.
- There are still a smaller number of uses where agentic AI is used inappropriately, such as by giving it tasks it is not capable of performing reliably. For example, 'shadow AI', where employees use agentic AI without permission, remains a problem. This is the case particularly where the person using it sees a benefit but not the wider risks. This leads to a small number of high-profile failures, compounded by the risk-taking approach of the developers that do use low-capability AI in high-risk situations.
- Our interaction with agentic AI occurs mainly through public complaints about bad experiences and data breaches. These cases involve organisations using low-capability agentic AI poorly or for tasks it is not suitable for, but the volume is not excessive.
- Analysts expect a reduction in investment in agentic AI. Technology companies continue internal research and development to improve capabilities, reliability and robustness. They continue to create

developer tools and **application programming interfaces** (APIs) to encourage experimentation, but they do not aggressively push agentic AI on mainstream users. Analysts expect a reduction in agentic AI hype. Reduced investment increases the cost of using agentic AI for the average person, further discouraging use. Open-source development continues for hobbyists and some specialised applications.

Scenario two: just good enough to be everywhere (high adoption, low agentic capability)

In this scenario, we see high adoption and use of agentic AI technologies despite the limited capabilities of agentic AI. Many of the problems with LLMs remain. This scenario may be driven by one or more of the following:

- Aggressive marketing and release of agentic AI.
- Low public and business awareness of flaws and friction around the implementation of agents.
- Assessments by users that the limited capability is 'good enough'.
- Users ignoring agents' limitations in the pressure to deploy agents.

Many of the harms in this scenario come from failures of agentic AI or ill-considered deployment. We would expect to see the following:

- Regular inappropriate or ill-advised use of agentic AI – agentic AI used where it is not technically suitable, including in high-risk areas (eg financial services, law or healthcare).
- High and low impact failures of AI agents occur regularly. Misinterpreted tasks, superficial approaches to tasks or failures on edge cases also cause frequent inconvenience.
- Low-capability agentic tools widely embedded in services (eg shopping applications, social media, banking portals, government services and education platforms). Conversational agents become a more common interface to these services. However, they are limited and often make mistakes.
- With large-scale adoption and agents mediating transactions, large volumes of personal information flow through agents and infrastructure providers.
- Processing of personal information by low-reliability agentic AI may lead to data protection compliance issues. These could include:
    - data breaches caused by sharing the wrong information;
    - collecting and processing personal information without a legal basis; or

    - creating security vulnerabilities.
- Agentic AI becomes an everyday part of our regulatory activity; however, its low capabilities lead to a high demand for intervention. Over time, we become very familiar with complaints related to data processing by agents.
- Frequent errors result from hallucinating or limited agents. There is a high requirement for human-in-the-loop oversight and constrained autonomy (task-by-task authorisation and frequent checkpoints). Supervision costs for checking and monitoring AI agents are high (although irresponsible or ill-informed users might skip them).
- Agentic AI providers use the requests, tasks, reasoning traces and outputs from these early agents to train more sophisticated models.
- The volume of AI agents puts pressure on online systems (automated access to websites, mass data scraping, frequent and excessive API calls), including online public services. There are chaotic and unpredictable interactions between different agents. These significantly reduce transparency, making it harder to uphold data rights.
- 'Correcting the agent' becomes a routine part of work and life, and this extends to data protection tasks. There is a high demand for redress and fixes for mistakes. Public dissatisfaction with agentic AI grows.
- Agentic AI decreases information security by introducing new types of unmitigated vulnerabilities to information systems.
- Large-scale generation of inaccurate personal information by agentic AI systems occurs.
- There is a proliferation of 'fake' agentic AI (other technology marketed as 'agentic') and low-paid human workers standing in for or covering AI agents. Agents may look more transformative than they are in practice.
- Demand for best practices for agentic AI deployment is high to minimise the harms.

Scenario three: Agents in waiting (low adoption, high agentic capability)

In this scenario, we see low adoption of agentic AI, despite increased agentic capability and overcoming many of the limitations that currently apply to LLMs and prototype agents. For example, agents include systems to mitigate LLM hallucinations and tools for meaningful governance and control of agent actions. Factors outside the technology itself could drive low uptake and use. These factors could include:

- increasing costs;

- time needed for changes in business models to fully take advantage of agent capabilities;
- residual public distrust;
- political barriers around access to agents or to computing infrastructure;
- lack of governance or liability frameworks;
- caution around sharing personal or confidential data; or
- focus on specialised applications where agents work well.

Here, data protection and privacy harms largely come from misuse of effective agentic AI, rather than errors. For example, issues arise from unwarranted intrusion or loss of control of personal information, rather than processing of inaccurate personal information. In this scenario, we would expect to see the following:

- A smaller number of agentic AI uses and a smaller demand for regulatory intervention. The number of agentic systems is also limited, reducing the impacts of complexity on transparency.
- Organisations deploying agentic AI tend to be knowledgeable and familiar with the technology's capabilities. This may mean the highest adoption in areas such as research labs and technology companies.
- Promising pilots of agentic AI happen but are not rolled out more broadly. Pilots and research projects using agentic AI still involve high-sensitivity data and use agents for complex, novel and data-intensive tasks. Developers may prioritise proof-of-concept over privacy by design.
- Organisations view the legal, compliance and reputational risks as outweighing any potential efficiency gains for wide deployment of agentic AI. This is particularly true in highly regulated sectors or high-stakes tasks.
- Agentic AI implementations are bespoke and customised for specific uses and users. No two implementations are the same.
- For the ICO, no two investigations or interventions around agentic AI are the same. When we have to investigate agentic AI uses, these are likely complex.
- Low public familiarity with agentic AI and relatively low use in everyday life. People are more familiar with early, frequently faulty AI. Users are mostly specialists. People's awareness of the capabilities of agentic AI is low. However, they may still be exposed to backend agentic AI by its specialist users.

- Agentic AI failures are relatively rare. Failures with a public interest get high media attention, given the background scepticism about AI.
- With the high capabilities of agents, employees may quietly and unofficially make use of agents without the permission, knowledge or support of their employers. This is a point of tension in these workplaces and causes data protection compliance errors and privacy harms.
- Technical ecosystems and the surrounding infrastructure for agentic AI (integration standards, best practices, oversight mechanisms, tools) are limited.
- Technology companies may increasingly shift effort from increasing raw capability of agentic AI to trust-building and risk mitigation efforts. Sophisticated agentic AI tools are used within these organisations, potentially giving them a strong advantage over competitors.

Scenario four: Ubiquitous agents (high adoption, high agentic capability)

In this scenario, we see high adoption of agentic AI that has increased in capability, approaching the marketing hype of its early years. This is not artificial general intelligence (AGI), and AI agents are not sentient. But they are powerful and capable tools, more reliable and robust than today. In this context, data protection harms arise from agents 'working as intended'. They still have impacts on people, such as privacy-invasive agents, or from inexperienced or malicious users tasking agents to build software that violates data protection requirements. In this scenario, we would expect to see the following:

- Large numbers of effective AI agents, widely deployed across many different parts of the economy and society. A wide choice of AI agents is available, with marketplaces for agents and 'skills' similar to app stores.
- Agents regularly access and process personal information. Because adoption is widespread, agents mediate massive amounts of personal information (including special category data) and organisational information.
- Agentic AI becomes an everyday part of our regulatory activity. Agentic capabilities integrate into many of the data processing activities we work with or oversee.
- Similarly, our investigations become highly complex because of agentic AI. When mistakes occur, they are hard to spot among the volume of agentic activity.

- Agent-to-agent communication, including the exchange of personal information, becomes common (eg your shopping agent negotiating with a retailer’s sales agent). This creates a complicated flow of personal information.
- People lose privacy because they can easily instruct agents to search for and collate information from multiple sources. Generative and agentic AI tools make coding and building data processing systems easy and accessible to non-professionals.
- New business models based around agentic AI emerge.
- Other modes of user interfaces (such as voice) become more common. Interaction with software shifts from menus or dashboards to natural language. Productivity suites centre around asking an agent rather than manual workflows.
- Agents technically capable of performing compliance and governance tasks emerge, including those associated with data protection or freedom of information.
- Agent vendors are well-resourced and influential. Agent ecosystems emerge around the dominant players, potentially raising competition issues.

Next steps

Engagement, guidance development and collaboration

This report reflects our early-stage thinking on speculative opportunities and risks. You should not read it as ICO guidance or an explanation of our regulatory expectations on agentic AI. We welcome contact from any stakeholders wishing to continue the conversation. We encourage organisations that want to contribute to our future thinking on agentic AI to contact us at: emergingtechnology@ico.org.uk.

We commit to holding workshops with industry to gather further information on agentic AI, including on agentic capabilities and adoption, and how industry is mitigating data protection and privacy risks.

AI continues to be a priority for us. We have several strands of ongoing work set out in our AI and biometrics strategy with implications for agentic AI.

We are developing a statutory code on AI and ADM with implications for agentic AI. If organisations have evidence relating to agentic AI they wish to submit for our consideration in the context of developing the code, they can contact ai@ico.org.uk.

We will start our process to update guidance on ADM and profiling, in light of the Data (Use and Access) Act, with public consultations in 2026.

We will continue to work with stakeholders to further our understanding of agentic AI and to promote data protection by design and default in the development of agentic technologies. We will continue to monitor the market for new developments, identify key stakeholders and take action where we have concerns about data protection compliance.

We will continue to work with innovators developing agentic AI products and services. We encourage stakeholders working on agentic AI applications to access our free innovation support services. For organisations that are in the process of developing innovative products and services using personal information and agentic AI in the public interest, we would encourage them to explore our Regulatory Sandbox.

Digital Regulation Cooperation Forum (DRCF)

Working with partner regulators through the DRCF, we will continue to further our understanding of the effects and impacts of agentic AI in both the workplace and the home.

The DRCF has recently announced the launch of a Thematic Innovation Hub offering tailored engagement and regulatory advice on priority topics, the first focus of which will be agentic AI. The intent is to develop our collective understanding of how one another's regulatory regimes might apply to AI, and work to identify and resolve any points of conflict.

The DRCF's Horizon scanning and emerging technology work anticipates developments in technology and their regulatory implications. The workstream helps regulators proactively anticipate the changing technological landscape and gives the industry early awareness of regulatory implications relevant to their innovations. This year, the DRCF will consider agentic AI to understand how it may develop and the implications for regulators and industry. The DRCF will produce a public output in 2026.

International engagement

During 2025, we introduced the theme of agentic AI into discussions at the G7 Emerging Technologies Working Group and welcomed the sharing of experience and expertise in this emerging area. We will continue engaging with our international counterparts to identify collaborative opportunities to further discuss and consider the data protection implications, risks and opportunities posed by agentic AI.

Annex I: Methodology

In this annex, we provide additional information about the methodology we followed to produce this report.

Futurecast

We used a bespoke version of the UN Global Pulse Futurecast to project plausible futures and consider how the increased use of agentic systems might impact data protection for stakeholders. Participants at two internal workshops were assigned the roles of interested parties who would be affected by, or can shape the development of, agentic AI. As the future timeline developed, we generated events which would impact agentic AI.

We used the PESTLE framework as the basis for analysis and discussion. The framework encourages participants to identify sets of political, economic, social, technological, legal and environmental factors impacting the future. We repeated the simulation with different attendees to identify any patterns or recurring themes.

Some repeating themes emerged, including issues around control, transparency and economic impacts. Participants also raised considerations around regulation and how society might change as a result of widespread use of agentic systems.

Stakeholder engagement

Following our desk research phase, we progressed to engaging with stakeholders whom we expected to:

- influence the use of agentic AI in future;
- help shape the technological development of it; or
- have an interest in the issues (including social, privacy or security) arising from its use.

We conducted interviews with academics, interest groups and industry. These interviews helped inform our thinking on developing trends, potential use cases and critical uncertainties. They also allowed us to validate some of the conclusions drawn from the research phase and consider practicalities that organisations and people are addressing in the real world.

We asked stakeholders questions focusing on:

- technical development, use cases and the state of the art;
- best and worst case outcomes for agentic AI;
- information rights issues;
- control and governance mechanisms and risk mitigation for agentic AI;
- development of the market for agentic AI; and
- crucial actions for developing agentic AI to achieve positive outcomes.

We collated and consolidated stakeholder answers as another input for this paper and for developing our futures scenarios.

Scenario planning

For this topic, we adopted a scenario planning methodology to manage the uncertainty around agentic AI. Scenario planning is a foresight methodology to support organisational planning. The aim is to identify a range of possible futures, describe them and support flexible planning that will work across multiple futures. This allows teams to identify key uncertainties and create indicators to support decision makers in understanding the scenario they are heading towards.

Rather than trying to predict the future in the face of high uncertainty, the method creates multiple scenarios. This process helped us to be explicit about our assumptions. The method can scale up or down depending upon resources, but is best when iterative, expert-led and stakeholder-centred.

An early step is identifying potential variables for a set of scenarios. These should be truly variable (with both ends being plausible and clear). They should also be significant (relevant to important aspects of the technology and the role of the regulator, and likely to have a considerable impact). In addition, the variables should be independent, avoiding strong correlations between them. When combined, the variables should result in four or so distinct, plausible and robust scenarios. It is always necessary to select variables, so the output of the scenarios needs to do something useful for the decision maker.

Based on our research and stakeholder engagement, we compiled a longlist of potential variables that could be pivotal for the development of agentic AI. These included the following:

- **Various ways of expressing or evaluating the capabilities of agentic AI:**
    - Autonomy or agency of agents (from narrow, task-specific tools to being capable of fully autonomous decision-making)
    - Generality of agentic AI (from dedicated agents for single tasks to general agents potentially adaptable to many or any tasks)
    - Ability of agentic AI to handle underspecification
    - Transparency and explainability of agentic AI
    - Goal complexity that AI agents can handle
    - Controllability or alignment of agentic AI with values or laws
    - Reliability of agentic systems
    - Efficacy of agentic systems
    - Security of agentic systems
    - Extent of the integration of privacy-enhancing technologies into agents
    - Extent to which limitations of LLMs have been overcome
- **Measures of the diffusion or adoption of agentic AI:**
    - Rate or extent of adoption/uptake of agentic AI
    - Accessibility of agentic AI to business or the general public
    - Rate of technological advancements in agentic AI
- **Dominant methods of use of agentic AI:**
    - Single agent for everything, multiple agents or integration of agentic AI into existing systems
    - Automated problem solvers to teamwork assistants
- **Market concentration or diversity:**
    - Open-source agentic models and tools or proprietary
    - Distributed or concentrated agentic innovation ecosystems
- **Social or political variables:**
    - Public trust in AI or in agentic AI specifically
    - Global regulatory alignment or fragmentation
    - Incidence of AI-related crises or scandals
    - Existence or strength of ethical norms around agentic AI use
    - Presence or absence of governance structures and mechanisms
    - Extent of liability mechanisms
    - Levels of AI literacy
    - Levels of AI governance

Steps taken to build and validate scenarios

We prioritised this long list based on:

- Impact – how significantly the variable could shape the future
- Uncertainty – how unpredictable the variable is
- Relevance – how closely it relates to our focal areas (in our case, information rights and our remit)

From this evaluation, participants in our internal workshop reached consensus around three top potential variables:

- AI agent adoption
- Market diversity
- Capabilities of AI agents

We then developed sets of four scenarios for each of the potential combinations of these three candidate variables. This allowed us to check for plausibility, to avoid overly correlated variables and to better refine our understanding of the extent of the variables.

We shared these mock-ups with key stakeholders internally and used them to frame questions to external stakeholders. We compared the sets of scenarios against each other to assess the type and quality of the insights they were creating.

We chose the extent of adoption of agentic AI and the general capabilities of agentic AI. We then returned to the findings from our research, stakeholder engagement and the Futurecast work to further populate and detail the four selected scenarios that we set out in this report.

To develop the scenarios in detail, we used a consistent set of indicators. This would help us to be consistent and make sure that each scenario was talking about similar things. This included asking the following questions, given the parameters of the scenario:

- What are key players doing in these scenarios?
- What are the potential data protection and privacy harms, and are there particular ways that these harms would manifest?
- What are the likely areas for innovation in agentic AI, and what is driving this?
- Where are the sources of economic growth?
- What are the business models and distribution channels around agentic AI?
- How is agentic AI being marketed or reported on?
- Are the public using agentic AI or agents?

- What do agentic ecosystems look like?
- What tasks do organisations and people use agentic AI for (either successfully or unsuccessfully)?
- What do agentic user interfaces look like?
- What are the opportunities for privacy by design?
- What cases and complaints might the ICO be seeing?

Annex II: Some drivers impacting the use of agentic AI

In our scenarios section, we briefly identify some of the drivers behind the emergence of agentic AI. In this annex, we provide more detail on these drivers.

Agentic AI drivers

Model training costs

Analysts project that keeping up with AI applications will demand compute power requiring USD $5.2 trillion investment by 2030. [34] The cost of training large AI models has escalated over time, with researchers predicting that by 2027 they will require billion-dollar investments. This could potentially concentrate power and control into only the larger technology providers who can invest at that level. [35]

A drop in compute prices and increased processing power driving accessibility

Research from Epoch.ai shows that the amount of physical compute required to achieve a given performance in computer vision models is decreasing at a rate of three times per year. This is driven by efficiency gains and systemic improvements. [36] There have been improvements in factors such as computational performance, compute required to achieve a given performance and the cost of training frontier AI models. Compute budgets are effectively doubled by the introduction of better algorithms every nine months. [37]

Increasingly large, high-quality datasets are available

Training datasets have grown over time, driven by demand and improvements in the creation and use of synthetic datasets. [38]

Venture capital funding and the AI bubble

In the first half of 2025, agentic AI startups worldwide received approximately USD $2.8 billion of venture capital funding. [39] Analysts predict that 10% of all AI funding in 2025 will focus on agentic applications.

In the UK, while the number of AI investment deals has decreased slightly, the average deal size has increased. Analysts expect this increase to create 6,500 jobs. [40]

'Fear of missing out' driven by the hype cycle and marketing

Gartner reports that organisations are ignoring the real cost and complexity of deploying agentic systems. This could ultimately lead to misapplication of the technology, driven by hype. They predict that over 40% of agentic AI projects will be cancelled by 2027 because of unclear business value and inadequate risk controls. [41]

Further contributing to the issue is the practice of 'agent washing'. This refers to vendors rebranding other technologies as agentic to take advantage of interest in the technology.

Highly intersectional technology

One in six UK organisations is using at least one type of AI in the workplace. Performance increases in compute power, data and algorithms used to process that data to complete tasks, as well as greater awareness of AI applicability, are driving this. [42]

Previous government information shows the IT and telecommunications sector has the highest AI adoption rate at 29.5%, with the legal sector at 29.2% and hospitality, health and retail sectors at around 11.5%. [43] The majority of use is data management and analysis (9%). Other applications included natural language processing and generation (8%) and computer vision and image processing and generation (5%). This indicates a range of applications for analysis as well as content generation. People increasingly interact with AI at work, and on the devices and in the applications they use at home.

Cost savings from reduced staff costs and labour

While technologies such as generative AI have seen widespread adoption, they seem to have had little impact in terms of value generation. [44] Some argue that agentic workflow automation will be far more impactful, bringing personalisation, adaptability and resilience to operations. [45] Proposed frameworks for measuring return on investment in the implementation of agentic AI look at productivity gains, cost savings and increased customer satisfaction derived from ADM.

A push on AI from national governments

In January 2025, the UK government released its AI Opportunities Action Plan, which detailed its commitment to using AI for the economic benefit of the nation and the social benefit of citizens. The USA followed in July 2025 with its own AI Action Plan. In August 2025, China issued a guideline on its AI Plus initiative with the aim that the 'intelligent economy' will become a significant growth driver for the Chinese economy. All the action plans share a focus on innovation and growth.

34 McKinsey article on the cost of compute power

35 Time article on the cost of Building AI

36 Epoch AI article on Machine Learning Trends

37 Epoch AI article on Revisiting Algorithmic Progress

38 Epoch AI article on Trends in Training Dataset Sizes

39 Prosus report about the rise of AI agents in the workplace pg.14

40 UK Government Artificial Intelligence sector study 2024

41 Gartner article on predicted Agentic AI project failure rate

42 Office for National Statistics report on understanding AI uptake and sentiment

43 Capital Economics report on AI Activity in UK Business

44 McKinsey report on the state of AI

45 McKinsey article on Agentic AI advantage

Annex III: Glossary of terms

- **Application programming interface (API)** – An API is a set of rules or protocols that enables software applications to communicate with each other to exchange data, functions and features.
- **Artificial general intelligence (AGI)** – AGI is a hypothetical stage in the development of artificial intelligence in which a system can match or exceed the cognitive abilities of human beings across any task. To an extent, it means the replication of human intelligence in an artificial form.
- **Chain of thought (CoT)** – CoT is a method of engineering prompts of LLMs, particularly for complex tasks in which the AI will need to complete multiple sub-tasks. This technique aims to facilitate problem-solving by guiding the model through a step-by-step reasoning process by using a sequence of coherent steps.
- **Contexts** – Context computing is the term for the ability of a system or application to understand and engage with the situational context of its user. Contexts could include a user's location, local time, environmental conditions or preferences.
- **Control mechanisms** – Control mechanisms are any means of ensuring that an AI system performs in line with expected functionality. It sometimes refers to controls on AI systems that are not based on model training alone.
- **Data poisoning** – Data poisoning refers to manipulation of the data that a model is trained on or provided with in order to introduce vulnerabilities, compromises or biases. This can lead to reduced security or performance, or behavioural issues, resulting in harms.
- **Edge cases** – An edge case is a problem or situation that falls outside normal operation in software development. Edge case testing can expose unusual behaviour or flaws.
- **Generative AI** – Generative AI can synthesise new content (music, images, text, audio and video) from training data. Often trained on extensive datasets, these tools can exhibit a broad range of general-purpose capabilities.
- **Guardrails** – Guardrails constrain AI systems to ensure or prevent certain outputs. They might include tools that monitor the processing of personal information, or identification of banned content, for example.
- **Interfaces** – An interface is a point of interaction between two systems or components that allows them to communicate.

- **Multimodal** – Multimodal interfaces are systems that process combined user input modes (eg voice, gaze, gestures and movements) in a coordinated manner.
- **Privacy-enhancing technologies (PETs)** – PETs are technologies that embody fundamental data protection principles by minimising personal information use, maximising information security or empowering people.
- **Retrieval-augmented generation (RAG)** – RAG is the process of optimising the output of an LLM so that it references an authoritative source outside its training data, such as a database, before it generates a response.

Annex IV: Further reading

Further reading

- AI Security Institute – How to evaluate control measures for AI agents?
- Chan et al. – Harms from Increasingly Agentic Algorithmic Systems
- Desai & Riedl – Responsible AI Agents
- Future of Privacy Forum – Minding mindful machines: AI agents and data protection considerations
- Harvard Business Review – Organizations aren't ready for the risks of agentic AI
- MIT Technology Review – Don't let the hype about AI agents get ahead of the reality
- Narayanan & Kapoor – AI as Normal Technology
- NIST – Tool use in agent systems
- OWASP – Agentic AI – Threats and mitigations

Annex V: Acknowledgements

With special thanks to the following for their contribution to our research:

- Professor Ali Hessami, The Agentic AI Safety Community of Practice
- Nell Watson, The Agentic AI Safety Community of Practice
- Dr Vasilios Mavroudis, AI for Cyber Defence Research Centre, Alan Turing Institute
- David Mlcoch, Asteroid
- Patricia Shaw, Beyond Reach Consulting Ltd
- Clifford Chance LLP
- DeepFlow
- The Future of Privacy Forum
- Google
- Meta
- OneTrust
- Pydantic
- techUK
- Dr Adel Bibi, University of Oxford
- We and AI

We are also grateful to all others who contributed to our research.