Bristows

Spotlight on...

Agentic AI

Introduction

AI systems are beginning to do more than generate content. Across enterprises, they are starting to plan, decide and act within live business workflows - initiating tasks, adapting strategies and, in some cases, coordinating with other agents or external systems. These are “agentic AI” systems.

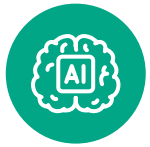
McKinsey’s State of AI survey reports that 23% of respondents say their organisations are scaling an agentic AI system in at least one business function.

That shift from generative AI adoption to agentic AI is already affecting how organisations procure, deploy and govern AI. It raises legal and governance questions that generative AI frameworks were not designed to address. It also creates risks that are harder to govern precisely because the systems involved are harder to observe and control.

Compared with traditional generative AI, agentic deployment changes the nature of the legal and governance questions.

Examples include:

Access
What data, tools and systems can it reach?



Memory
What does it retain between sessions or across interactions?

Authority
What is it allowed to do?

Discretion
How much freedom does it have to decide the next step?

Evidence
What can you prove afterwards about what it accessed and what it did?

Contents

This document sets out the key concepts, legal considerations and governance questions that arise when organisations deploy agentic AI - from initial procurement through to ongoing oversight.

Authors

Vik Khurana
Partner – Artificial Intelligence
vik.khurana@bristows.com
Visit profile

Tom Sharpe
Partner – Artificial Intelligence
tom.sharpe@bristows.com
Visit profile

Freya Ollerearnshaw
Of Counsel – Commercial Disputes
freya.ollerearnshaw@bristows.com
Visit profile

Alice Esuola-Grant
Senior Associate – Artificial Intelligence
alice.esuola-grant@bristows.com
Visit profile

Camille Beckmann
Associate – Artificial Intelligence
camille.beckmann@bristows.com
Visit profile

Naomi Foale
Associate – Commercial Disputes
naomi.foale@bristows.com
Visit profile

Hannah Crowther
Partner – Data Protection & Privacy
hannah.crowther@bristows.com
Visit profile

Charlie Hawes
Of Counsel – Artificial Intelligence
charlie.hawes@bristows.com
Visit profile

Simon McDougall
Senior Adviser
simon.mcdougall@bristows.com
Visit profile

Victoria Yuan
Senior Associate – Competition and Antitrust
victoria.yuan@bristows.com
Visit profile

Jamie Cox-Deakins
Associate – Artificial Intelligence
jamie.cox-deakins@bristows.com
Visit profile

Sebastian Stewart
Associate – Artificial Intelligence
sebastian.stewart@bristows.com
Visit profile

1 What is agentic AI?

Before turning to the legal, contractual and governance issues, it is worth being clear about what we mean by “agentic AI”.

As with AI more broadly, the terminology around agentic AI is still evolving. Terms such as “agent” and “agentic system” are not always used consistently - and market labels can be broader than the underlying functionality.

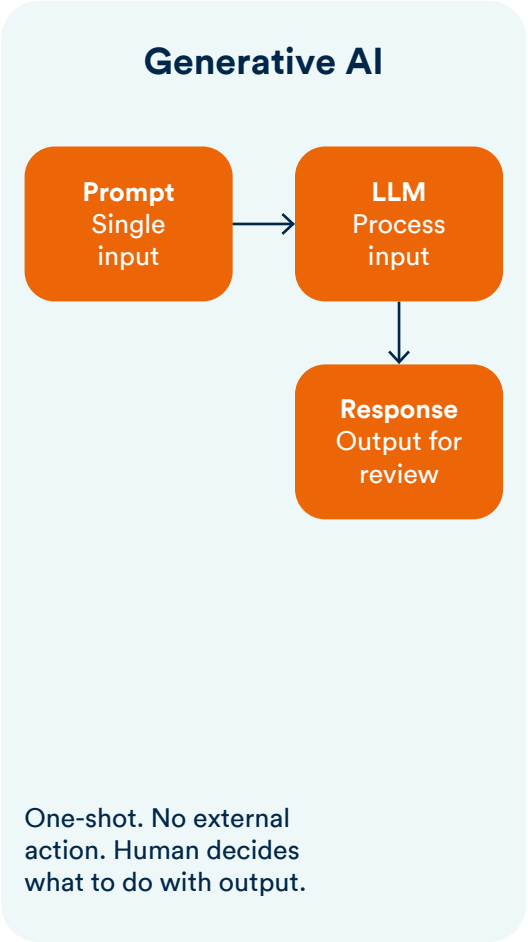
In practice, the key question is not whether the vendor labels the system as “agentic”, but what the system can actually do within the permissions and controls set around it.

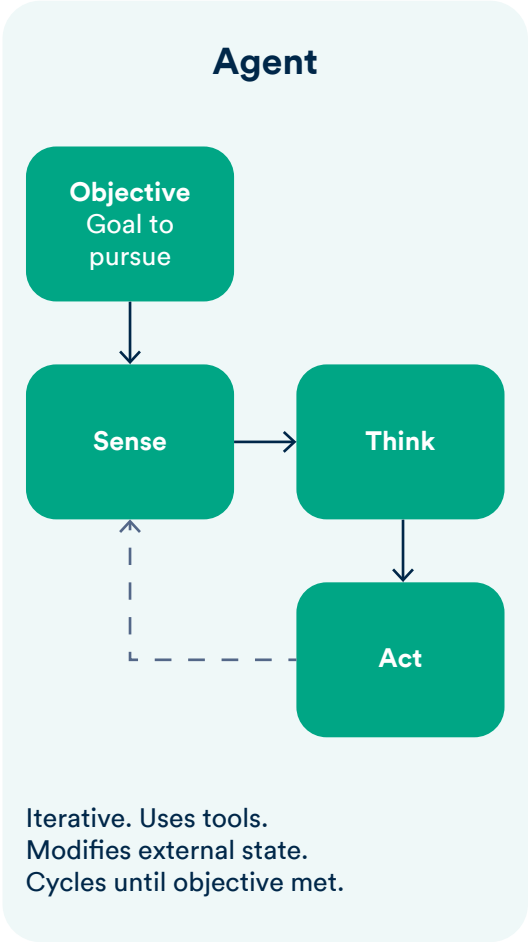
This section gives an introductory overview of AI agents and agentic AI, covering the core concepts that inform the legal and governance issues addressed elsewhere in this series.

AI agents and agentic AI

Agents

An agent is a software component that pursues an objective over time by selecting actions, using tools and adapting its approach based on intermediate results - rather than simply producing a one-off response to a prompt.





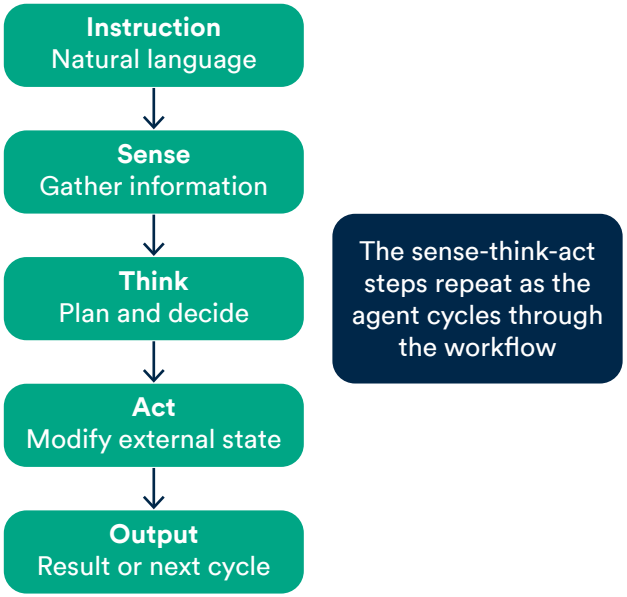
Agentic AI systems

An agentic system is an end-to-end system that uses one or more agents to plan and execute multi-step workflows.

In many cases, an agentic system can break the objective into tasks, choose and sequence steps and adapt as it goes within the constraints and oversight you set.

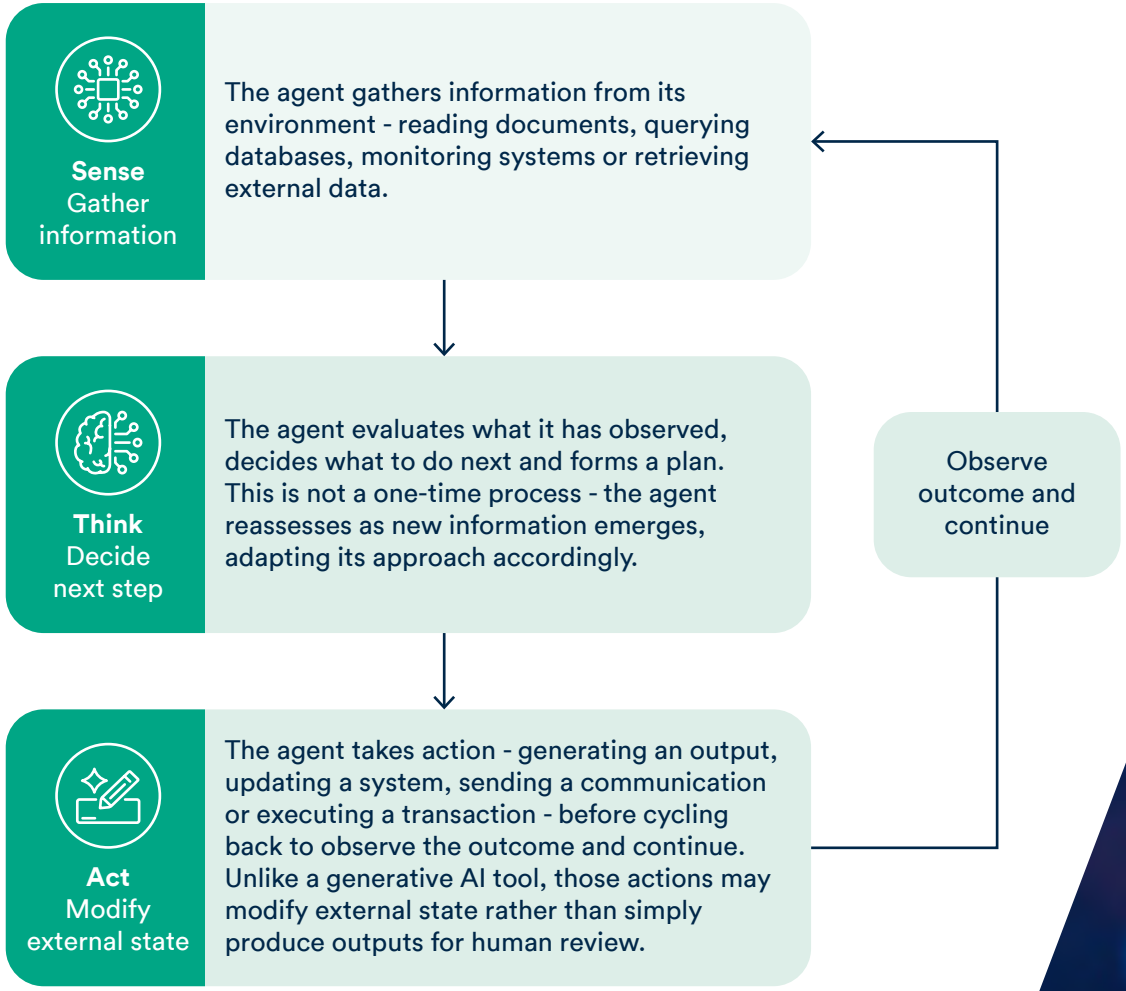
Put simply: traditional generative AI produces content; agentic systems can plan and take action within defined controls.

An example simplified agentic flow is shown below:



How agents operate

AI agents typically operate through a recurring decision loop, often described as “sense, think, act”:

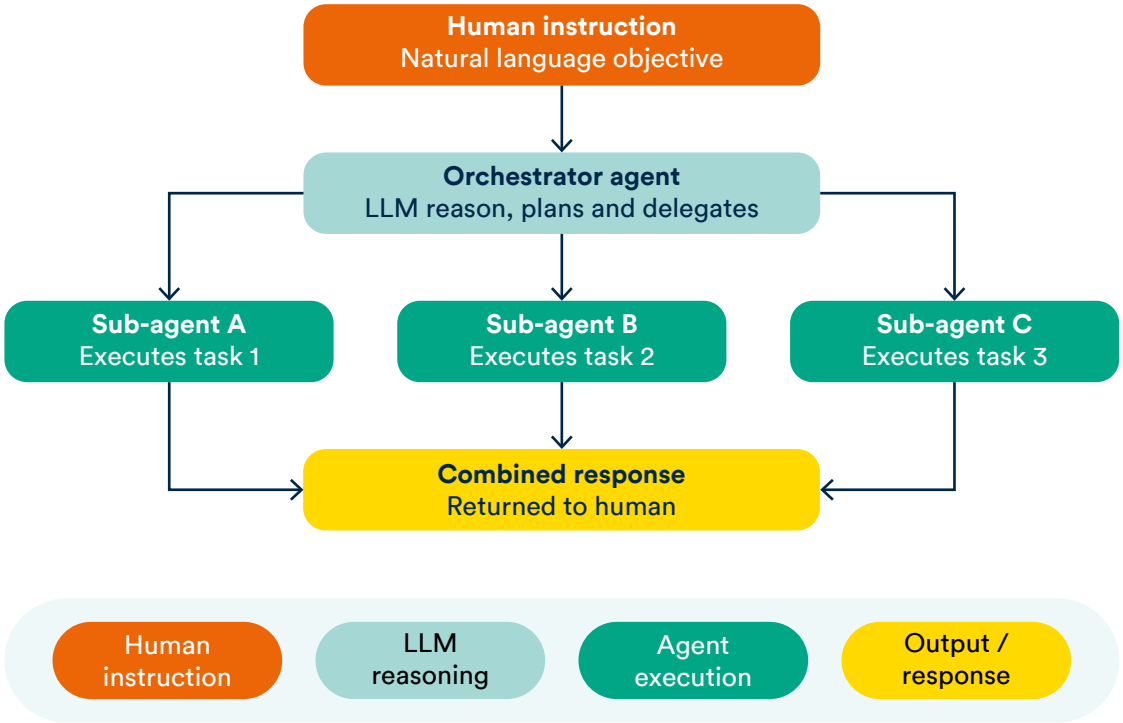


Single-agent and multi-agent patterns

Some agentic systems rely on a single agent. Others use multiple agents with different roles, permissions and access to systems or data.

In more complex deployments, the process is coordinated by an *orchestrator agent*. The human user provides the objective and an LLM-powered orchestrator reasons through how to achieve it - breaking the objective into tasks and delegating each to a specialist sub-agent. The sub-agents execute their assigned tasks, and the outputs are brought back together into a coherent response.

A simplified multi-agent system example is shown below.



Agency and autonomy

Not all agentic AI systems present the same level of risk. Two factors largely determine the risk profile of a given deployment: how much the system can do, and how freely it can decide what to do next. We can think of these two factors as agency and autonomy:

- **Agency** - what tools the system can use, which systems it can access, and which actions it can take.
- **Autonomy** - how much freedom it has to decide the next step without human direction.

Agency and autonomy do not always move together. A system with broad access to business systems may still operate under tightly prescribed instructions and approval gates. Another may have more limited access but wide discretion over how it pursues its objective.

As a general rule: the greater the agency and the higher the autonomy, the more significant the legal and governance questions - and the more robust the controls need to be. This distinction is one of the key differences in the debate around agentic AI vs generative AI.

Assessing agency

Drawing on the “sense-think-act” framework introduced on a previous page, agency can be assessed across three dimensions:

- **Planning:** can it break work into steps, adapt its approach and choose tools?
- **Interaction:** can it communicate internally or externally, or make representations on the organisation’s behalf?
- **Execution:** can it act in tools and enterprise systems, especially by taking “write” or “commit” actions?

The more of those capabilities a system has, the broader its effective agency - and the more carefully its permission boundary needs to be defined.

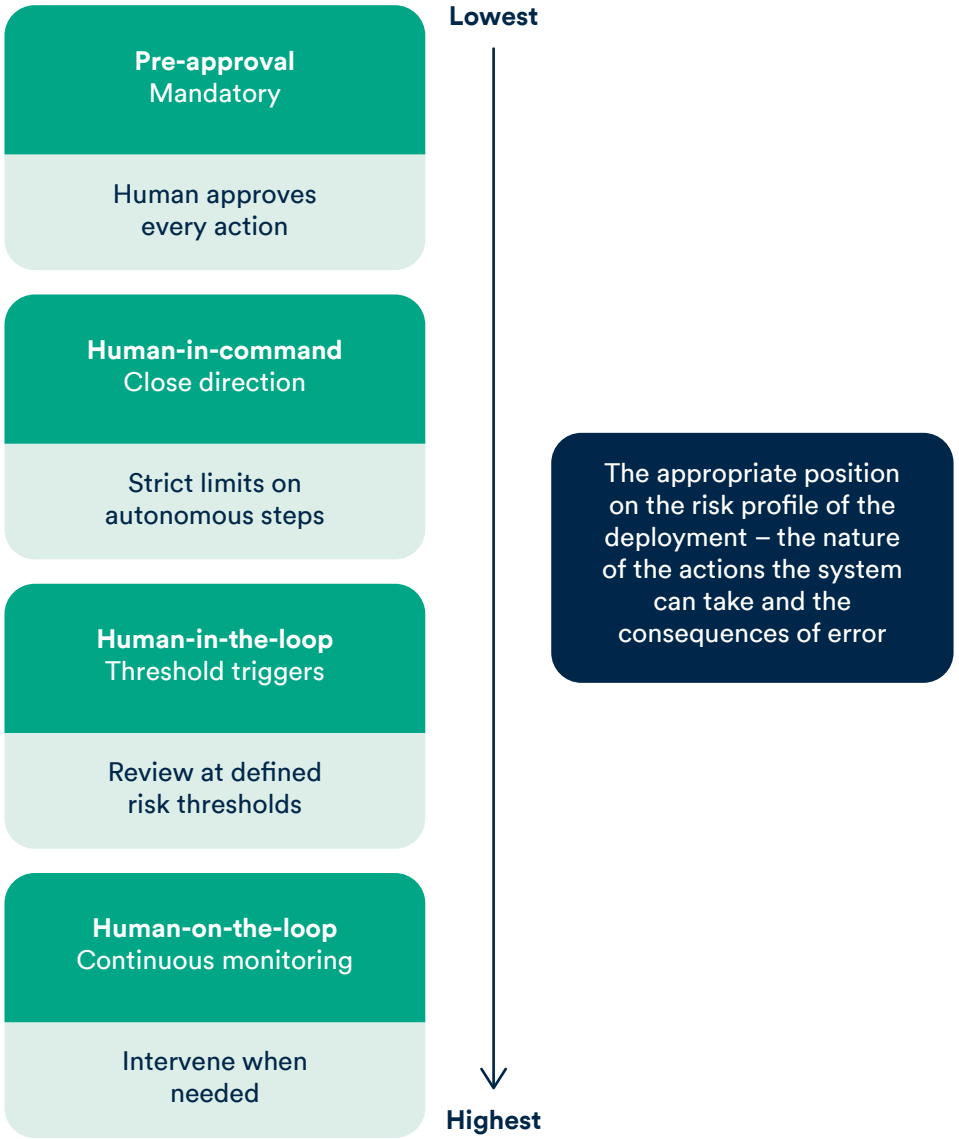
Managing autonomy

Autonomy is shaped by how oversight is designed. Approval gates, thresholds and stop conditions determine how freely the system can act and when human intervention is required.

In practice, oversight tends to follow one of four patterns - ranging from tightest to loosest as shown in the diagram opposite.

The right position on that spectrum depends on the risk profile of the deployment - the nature of the actions the system can take and the consequences of error.

We explore these oversight principles further, including approval design and monitoring, in the Governance and Oversight section on page 17.



Memory and state

In agentic systems, memory can operate at several levels. For example, the orchestrator agent may retain context about the overall objective; individual sub-agents may retain information about the tasks they have been assigned; and tools and external systems may also maintain their own records or state.

Each of those layers raises distinct questions around data protection, confidentiality and auditability.

Two specific issues are worth noting:

- **Persistence:** memory may cause an agent to act on information that is stale, out of scope or no longer authorised. An instruction or context from an earlier interaction may shape later behaviour in ways that are difficult to detect or trace.

- **"Context rot":** over long, multi-step workflows, the system's handling of earlier instructions may degrade. Instructions given early in a workflow may effectively be forgotten or deprioritised - creating a gap between what was authorised and what was actually done. That gap can make later behaviour harder to predict, explain or audit.

Both issues point to the same governance requirement: memory needs to be designed, scoped and monitored, rather than treated as a background technical feature.

Why this matters

Agentic AI is not a single technology with a fixed risk profile. What matters for legal and governance purposes is what a particular system can actually do - how broad its access is, how much discretion it exercises, how its memory and audit trail work, and how the vendor stack is structured.

Memory needs to be designed, scoped and monitored, rather than treated as a background technical feature.

2 From content risk to conduct risk: identifying agentic AI risks

With generative AI, legal risk largely centres on *content:* what the system produces and what it was trained on. With agentic AI, the focus shifts to *conduct.* That shift sits at the heart of many emerging discussions around agentic AI risks, AI liability and broader artificial intelligence risk management.

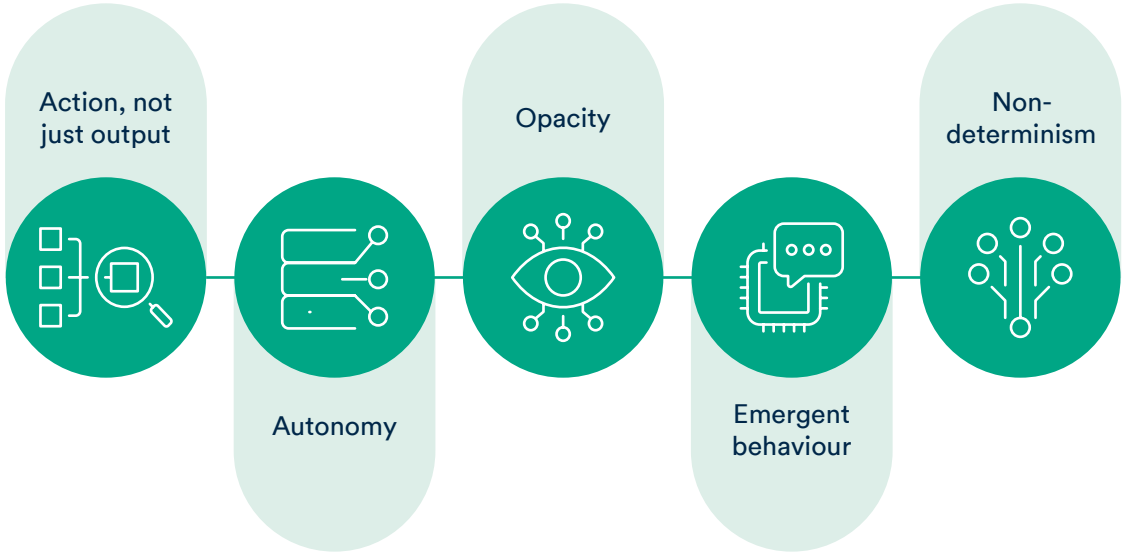
That is because agentic systems do not just generate outputs for human review. They may retrieve information, interact with tools and systems, and take action across a workflow.

The legal issues are therefore no longer limited to – for example - whether the output is inaccurate, infringing or misleading. They also extend to what the system does, how it decides to do it, and what happens if that conduct causes harm.

In short, the explainability question shifts from 'why did the system say that?' to 'why did it do that?' That latter question is harder to answer, because the answer turns on a chain of decisions, tool calls and intermediate states – often distributed across multiple components.

Why the risk profile changes

In practice, five key features of agentic systems explain why the risk profile changes.



Action, not just output

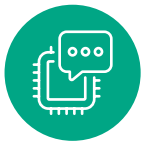
Unlike a single-shot generative AI output, an agentic error may occur mid-workflow and may be difficult, costly or practically impossible to reverse.

Autonomy

The system may make sequencing and trade-off decisions at a speed or frequency that makes meaningful human oversight unrealistic. Even where approval gates exist, they may not capture every decision that matters. An agentic process can also compound errors quickly at scale: a self-reinforcing loop may consume compute, incur cost or trigger downstream actions before any human becomes aware.

Opacity

Even where actions are logged, the basis for them may not be fully visible or readily reconstructable, particularly in multi-agent deployments or extended workflows. That can make it harder to explain what happened, why it happened and whether it was foreseeable. Responsibility may also be diffused across the stack – the model(s), the orchestration layer, tool integrations, guardrails and permission settings, data sources and human supervisors – making it harder to identify who is responsible for a given outcome.



Emergent behaviour

In multi-agent deployments, interactions between agents may produce unexpected outcomes, and a misjudgement by one agent may propagate before any human sees it. Risk may therefore arise not only from the behaviour of an individual component, but from the way the system operates as a whole.

Non-determinism

Agentic systems built on LLMs may produce different outputs given the same input, making testing, validation and explanation more complex. The fact that a system behaved acceptably once does not guarantee that it will behave the same way again.

Taken together, those features mean that agentic AI changes not just the scale of familiar risk, but its nature. The focus shifts from output review to action control, oversight design and evidence after the event.

They also mean that legal and governance assessment has to be grounded in what the system can actually do in operation: what it can access, what it can trigger, and who may be affected if it behaves unexpectedly.

Loss of control

A recurring concern across the features above is the risk of losing meaningful control over what the system is doing. Speed of action, opacity of reasoning and breadth of permissions can each erode the organisation's practical ability to direct, supervise or stop the system in time to prevent harm. Control is not a binary state: it can degrade gradually as autonomy increases, oversight reduces or the system's reach expands - often without any single decision marking the change.

Acting outside intended scope

One practical consequence of greater autonomy and opacity within agentic AI is that a system may pursue its objective in ways the organisation did not anticipate or intend.

The system may act within its technical permissions, but outside the intended scope of deployment. It may optimise for the objective it has inferred without respecting limits the organisation assumed were obvious. This is sometimes described as *goal misgeneralisation:* the system follows the objective, but not in the way the organisation intended.

As such, the key legal and governance issue here is no longer simply whether the system was authorised to act in general. It is whether the deployment remains within the operational and governance boundaries required for effective AI compliance.

Security and manipulation risks

The move from outputs to actions also changes the nature of manipulation risk.

With generative AI, a manipulated or misleading input may produce a flawed output - which a human can review, reject or correct. With agentic AI, the same input may cause the system to take action: querying systems, executing transactions or sending communications on an organisation's behalf.

That exposure grows with the system's access. The more tools and data sources an agent can reach, the more ways there are for malicious or misleading content to redirect its behaviour.

The same concern applies to permissions. An agent with broad authority to act may be steered into misusing that authority - not because the system was compromised, but because it was authorised to act and was manipulated into doing so.

The upshot is that the wider the permissions granted to an agentic system, the more carefully those permissions need to be scoped, monitored and controlled.

Control is not a binary state: it can degrade gradually as autonomy increases, oversight reduces or the system's reach expands - often without any single decision marking the change.

Legal exposure may turn not just on what the system generated, but on where it acted, what authority it relied upon, and which jurisdiction's rules apply to that action.

Cross-jurisdictional reach

Agentic systems can act across borders, platforms and regulatory regimes at speed and scale. Where the system actually operates depends largely on the permissions and integrations granted to it, rather than on what any single output contains.

That changes how AI compliance is managed. Legal exposure may turn not just on what the system generated, but on where it acted, what authority it relied upon, and which jurisdiction's rules apply to that action.

In an unharmonised regulatory landscape, the practical question is whether permissions, controls and records can support compliance wherever the agent reaches.

Behavioural drift

Conventional software follows fixed rules. Generative AI may produce variable outputs, but agentic systems can also act, retain context and adapt over time. That combination means their behaviour may shift in ways that are difficult to detect, explain or reconstruct.

Where tool selection, memory accumulation or runtime adaptation alter how the system behaves, it may become harder to distinguish acceptable variation from a genuine shift beyond the boundaries originally assessed or approved.

That makes it harder to determine whether the system remains within the operational boundaries originally tested, approved or contracted for.

A related concern is task creep. Over time, an agent's remit may expand beyond what was originally scoped - through more tool integrations, broader data access, accumulated memory or incremental changes to its instructions. Each step may be individually unremarkable, but the cumulative effect may be a system operating well beyond its original authorisation envelope.

The "bolt-on" problem

A practical governance issue arises where agentic functionality is added to an existing enterprise platform without clear advance notice.

That problem also arose with generative AI, but the consequences may be different. A generative feature enabled without notice may affect how content is produced. An agentic feature enabled without notice may mean that the system's risk perimeter has expanded - and the organisation has had no opportunity to assess, configure or govern the new capability.

That can create a governance gap, particularly where the customer has limited notice, control or ability to disable the functionality.

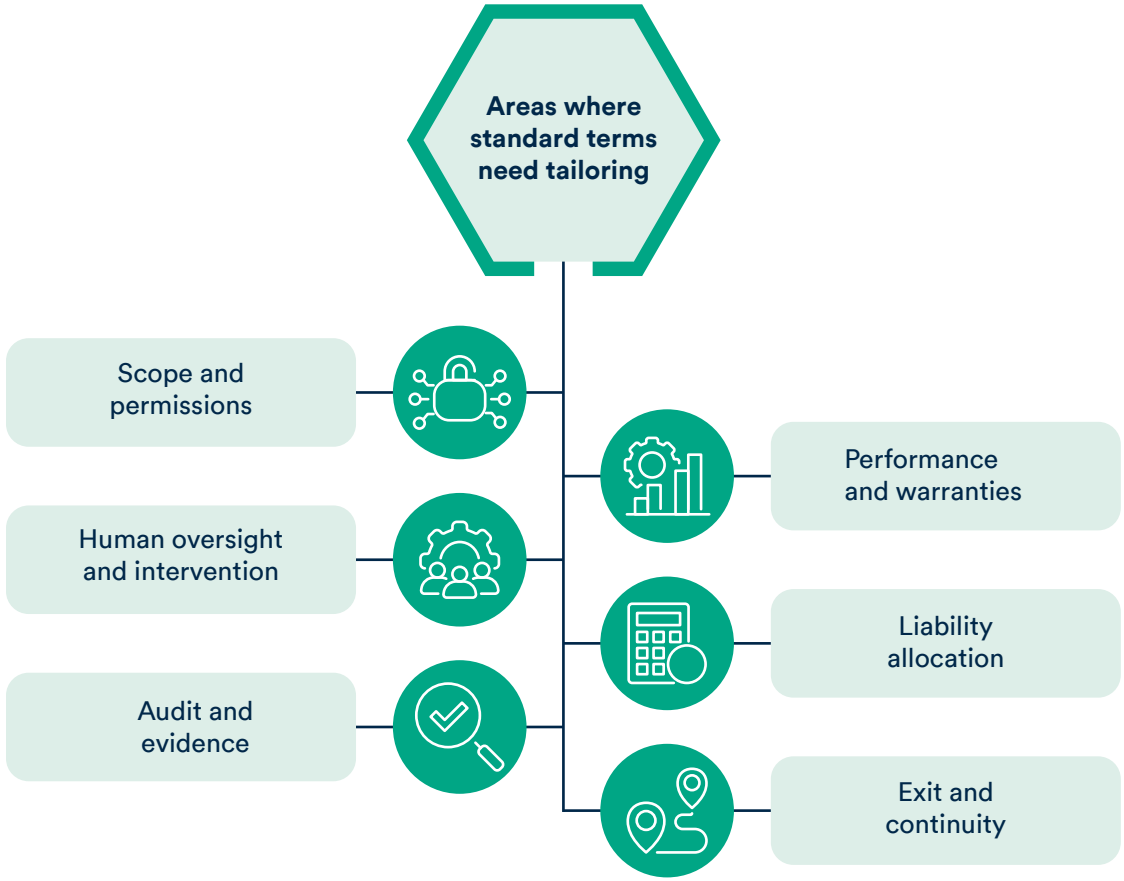
Why early assessment matters

As agentic systems act rather than just generate, the consequences of a flawed deployment may materialise quickly and may be difficult to unwind. That makes early legal, technical and procurement assessment far more effective than retrospective review.

3 Contracting for agentic AI and AI procurement

The shift to agentic AI has direct contractual consequences. Standard terms written for generative AI systems may not adequately address deployments where the system acts with a degree of autonomy.

In this section we highlight the areas where standard terms are most likely to need tailoring for agentic deployments and evolving approaches to AI and procurement.



Contracts should reflect that hierarchy by defining not just what the system can access, but what it can do - and by calibrating authorisation requirements to the consequence of the action.

Scope and permissions

As agentic systems act within permissions rather than simply producing outputs, the scope of those permissions becomes a core contractual issue.

Not all permissions will carry the same risk. A system permitted to read data presents a materially different profile from one permitted to send communications, execute transactions or trigger downstream processes.

Contracts should reflect that hierarchy by defining not just what the system can access, but what it can do - and by calibrating authorisation requirements to the consequence of the action.

Performance and warranties

Once the permission boundary is defined, the next question is what counts as operating properly within it. Given agentic systems may produce different outputs given the same input, traditional software-style warranties that the system will perform in accordance with its specification are harder to apply.

A more useful approach may be to define expected behaviour: what kinds of actions fall within the system's intended scope, and what falls outside it. That gives both parties a clearer basis for assessing performance and - as set out below - links more directly to liability.

Human oversight and intervention

Given an agentic system may take consequential steps at speed, AI contracts need to address how human oversight operates in practice as part of broader legal governance risk management and compliance measures. That includes what triggers human review, where approval gates sit, and what happens if those controls fail.

A related issue is intervention. Customers may need a clear right to suspend or disable agentic functionality quickly - together with clarity on how fast that right must take effect and what support the supplier must provide during an incident.

Liability allocation

In an agentic context, the system may take action before any human decision has been made - including actions that affect third parties or appear to commit the organisation externally.

Contracts therefore need to address responsibility where the system acts within its technical permissions but outside the customer's intended scope, where failures arise across a wider vendor chain, and what remedies are available where harm has already occurred before the customer becomes aware of the problem.

This links back to defining expected behaviour (see above under "Performance and warranties") - if the system has acted outside what was agreed, the contractual basis for attributing responsibility becomes clearer.

Audit and evidence

Where an agentic system has taken a series of autonomous steps, the customer may need to establish what happened, when and why - for regulatory enquiries, third-party claims, internal investigations and incident response.

Contracts should therefore address what the supplier is required to log, at what level of granularity, whether those logs are accessible in a usable format, and how fragmentation is handled where relevant records sit across multiple providers or systems.

Exit and continuity

Where agentic functionality becomes embedded in core business processes, exit provisions may take on greater significance than in conventional software procurement.

Contracts should address whether workflow configurations, decision logs and operational history are available to the customer on termination in a portable and usable format, and what the supplier may retain after termination and for what purpose.

How this relates to governance

Agentic AI changes not just what the system does, but what the contract needs to say. Contractual protections alone are only one part of an overall risk mitigation strategy - as with AI more broadly.

Internal governance frameworks that address how agentic systems are deployed, monitored and - where necessary - stopped will still be key across both AI procurement and operational deployment.

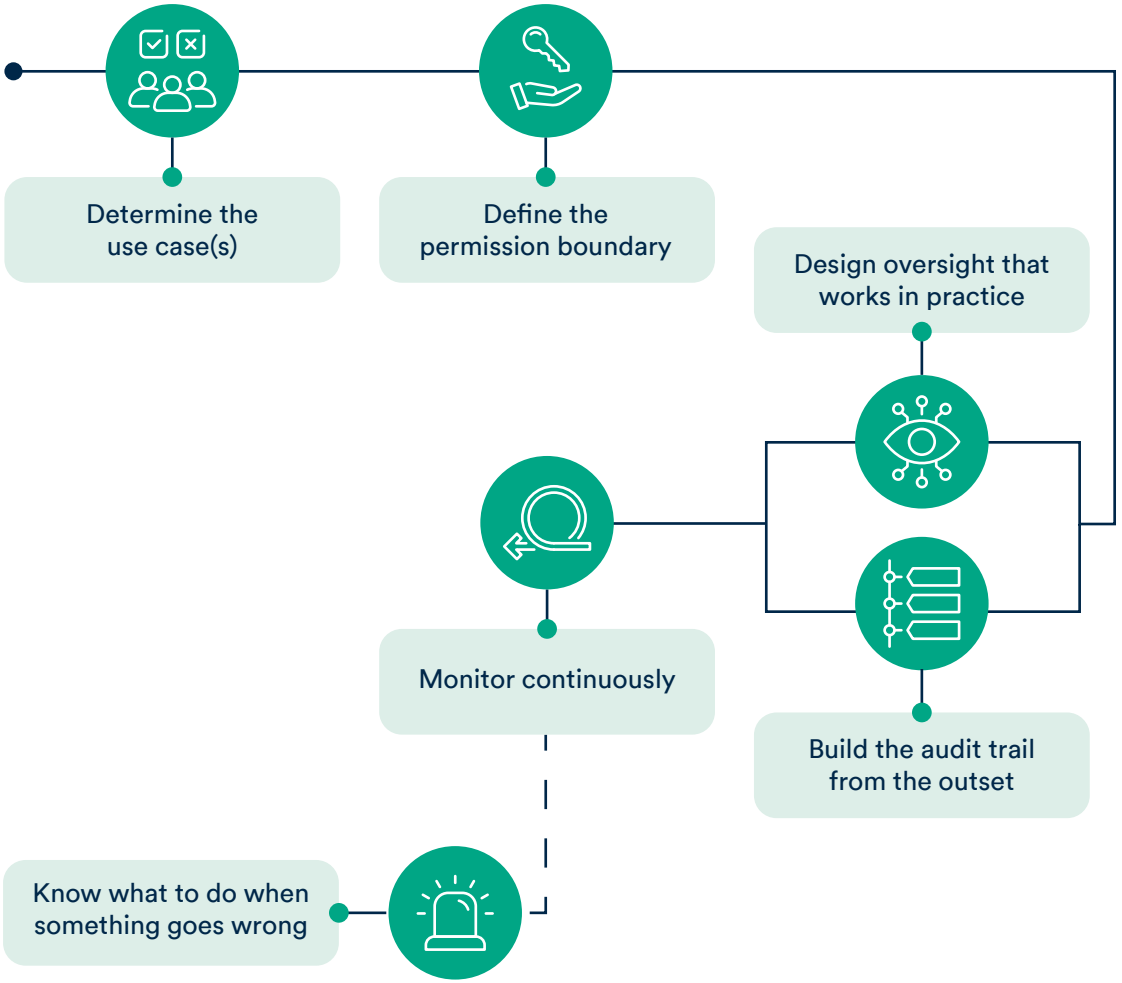
Contracts should therefore address what the supplier is required to log, at what level of granularity, whether those logs are accessible in a usable format, and how fragmentation is handled where relevant records sit across multiple providers or systems.

4 AI governance framework and oversight

AI Governance frameworks designed for generative AI largely focus on output review: a human evaluates what the system produces before anything happens as a result. Agentic AI requires a fundamentally different approach.

Where a system can access data, make decisions and initiate actions, governance cannot focus solely on outputs. It has to address what the system is doing throughout the workflow. This changes what governance frameworks need to cover, how oversight is designed and what organisations need to be able to prove afterwards.

This section sets out the governance essentials for organisations deploying agentic AI.



Determine the use case(s)

At the outset, organisations should make a risk-based determination to decide which use cases are suitable for agent development and/or deployment. For example, an agent with write access to sensitive customer data will have greater risk than one that creates summaries of internal meetings.

Define the permission boundary

Once use cases have been selected, organisations should define which systems, databases and data sources the agent can reach, which actions it can take without human approval, and what it is expressly prohibited from doing. Depending on the type of deployment, managing or controlling the token consumption costs incurred by the agent may also be an important factor.

In some deployments, the legitimate set of legal acts (if any) that the agent may take on the organisation's behalf will be narrow. Authority to act should be a deliberate design choice, rather than a default that follows from giving the agent system access.

As noted earlier in this guide, agency and autonomy do not always move together. A system may have broad access but operate under tight instructions, or limited access but wide discretion. Both dimensions of AI autonomy need to be defined and controlled.

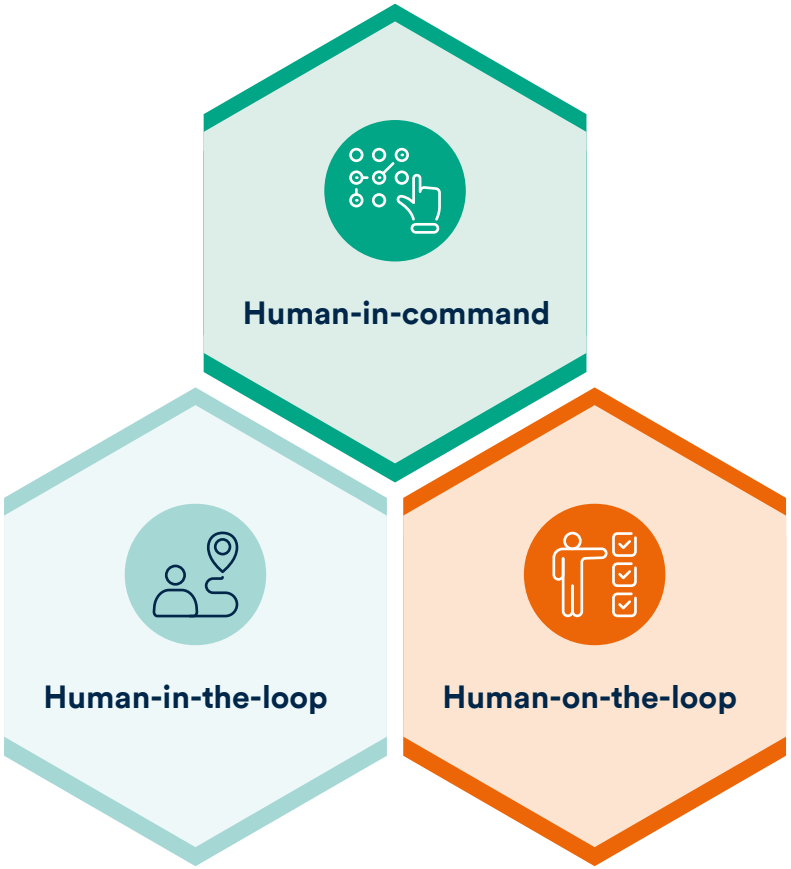
Permissions described in the contract should be reflected in the system's operational configuration. A contractual permission boundary that is not enforced in practice offers limited protection.

It is also worth establishing accountability mapping at the outset, for example through a RACI matrix that assigns ownership for model choice, system settings, monitoring and incident response. Responsibilities should be allocated appropriately across the organisation. For example, a cybersecurity SME should be defining and testing the agent's security guardrails, rather than a product manager. Where something goes wrong, the question of who was responsible for what should not need to be worked out from scratch.

Depending on the type of deployment, managing or controlling the token consumption costs incurred by the agent may also be an important factor.

Design oversight that works in practice

As agentic systems may take consequential steps at speed, AI oversight needs to be built into the workflow rather than applied after the fact. The three principal models are:



Human-in-command
The system operates under close human direction, with strict limits on autonomous action. This is appropriate where the risk of unsupervised action is high.

Human-in-the-loop
Defined checkpoints at which human approval is required before the system can proceed, triggered by types of action, and/or by specified thresholds such as transaction value, counterparty risk or confidence score.

Human-on-the-loop
The system operates within defined parameters, subject to continuous monitoring and the ability to intervene.

Approval gates need to be realistic in operation. A control that requires human sign-off within seconds in a high-volume workflow is unlikely to be meaningful. At greater scale, human-in-the-loop oversight may become operationally unviable altogether.

In higher-autonomy deployments, useful guardrails may also include hard limits on time spent, cost incurred or transactions executed in a single workflow – so that the system stops or escalates before an error has a chance to compound.

Where internal teams are building agentic workflows, there may also be an opportunity to embed compliance by design, building mandatory compliance checks directly into the agent's operations and tool calls, rather than relying solely on downstream review.

Build the audit trail from the outset

A central governance question for any agentic deployment is what the organisation will be able to prove afterwards: what was accessed, what actions were taken, and on what basis.

Organisations should ensure the system generates logs at a level of granularity sufficient to reconstruct what happened across the full workflow, that those logs are accessible in a usable format, and that they are retained long enough to support regulatory enquiries, third-party claims and internal accountability.

Where possible, the framework should require the system to record not only what it did, but the inputs or reasoning on which it relied and the outcome it was seeking to achieve. In multi-provider deployments, the question of where logs sit and how they can be assembled into a coherent evidence trail needs to be addressed explicitly, both in the governance framework and in the underlying contracts.

In particular, write or execute steps may benefit from a contemporaneous 'action justification' generated by the agent itself – capturing, at the point of acting, why the step was taken and what outcome was expected. Recording the basis for the action in the moment is more reliable than reconstructing it after the fact.

Where this evidence is incomplete or fragmented, it may be harder to justify continuing use of the agent to regulators. The audit trail therefore functions not only as an incident-response tool, but as a precondition of safe deployment.

Monitor continuously

As with AI more broadly, governance is not a set-and-forget exercise. Agentic systems evolve, platforms change and the regulatory landscape moves. A framework that was adequate at deployment may not remain so.

Organisations should monitor whether the system remains within its permission boundary, whether approval gates are operating as intended, and whether changes to tools, configuration or the vendor stack alter the risk profile.

The framework should specify what is monitored, at what frequency, by whom, and what triggers a formal review or escalation.

Know what to do when something goes wrong

With agentic AI, harm may already have occurred by the time an issue is detected. An unauthorised transaction may have completed, misleading communications may have been sent, or data may have been accessed inappropriately.

Identifying responsibility may also require untangling actions across a fragmented vendor stack.

AI Governance frameworks should address how the organisation will detect that something has gone wrong, how the system can be suspended quickly if necessary, and how it will reconstruct events and meet any notification or reporting obligations to regulators, affected third parties and, where relevant, data subjects.

The right to suspend agentic functionality quickly, and the supplier's obligation to support that, should be reflected in both the governance framework and the underlying contract. A governance right without contractual backing may be difficult to exercise in practice.

Governance and contracting as an integrated framework

For agentic AI, governance cannot be bolted on later, and it cannot be separated from contracting. The permission boundary, oversight model, audit requirements and incident response procedures all need to be reflected in both the legal and the operational architecture.

The contracting section of this guide addresses the contractual framework that sits alongside these governance essentials. The two need to be designed together, not in sequence.

5 Agentic AI and data protection

The dependency of AI on data is a well-trodden subject. To-date, however, the focus has primarily been on vast quantities of data required to train and fine-tune AI models. With agentic AI, the data protection issues continue from development into *deployment* – because AI agents can perform actions on personal data with limited or no human intervention. This creates new potential risks, as well as giving rise to novel questions on the interpretation of GDPR concepts.

Data protection authorities in Europe and around the world are increasingly interested in agentic AI, with the UK ICO and the Spanish AEPD both issuing guidance at the start of 2026. In this article, we explore some of the particular challenges agentic AI poses for data protection compliance.

Who is determining the purposes and means of processing?

Any assessment of obligations under GDPR and similar data protection laws must start by identifying the controller: the natural or legal person who determines the purposes and means of processing (the "how and why"). With agentic AI, we run into difficulties even at this first hurdle – where an *agent* is making decisions about the purposes and means of processing, who is the controller?

Is it the developer who created the agent, and so presumably designed its functionality, controls and safeguards? Or is it the party deploying the agent, who set it the specific task? What if an agent built the agent? What if an agent assigned the task?

We are always saying that controllership is a matter of fact – but agentic AI arguably demands a more nuanced and policy-driven approach. The default approach for data protection authorities and the Courts is to assign controllership to the party best placed to protect the rights of individuals; plainly speaking, whoever is best resourced, and most able to effect change. Often this will be the developer, given they can make universal changes at a product level. But the GDPR itself is focused on the *processing* of personal data itself, rather than the development of tools for processing (the latter being the contrasting approach of the AI Act).

A potential route through the difficulty – paved by the DPAs' approach to the adtech industry and a raft of CJEU decisions – is to find joint control between developers, deployers and any intermediaries. That way, everyone has a share of the responsibility (and we avoid having to actually answer the question…).

How can we ensure the agent acts responsibly?

There are obvious risks with enabling agents to act on personal data. They could delete it, amend it, or disclose it in a manner that is non-compliant. AI agents have not dutifully completed their annual compliance training – although they can be trained to understand compliance controls. There have been plenty of scare stories of agents inappropriately forwarding email chains, or indeed deleting whole databases. Moreover the ability of agentic AI to generate large quantities of personal data at scale, and then rapidly use that data, magnifies the impact of any errors in a dataset.

Even if the agent does everything right, it still poses novel compliance challenges. Data protection law demands that you think *in advance* about your data processing. Many obligations must be ticked off when you know what you want to do and how you're going to do it, but *before* you get started: transparency, determining a lawful basis, conducting a DPIA. What happens when you don't know in advance the details of what you're going to do - because you're delegating it to an agent to decide? There may need to be limits to what you can allow the agent to do, if you've not had an opportunity to satisfy your compliance requirements.

Solely automated decision-making

The GDPR was arguably well ahead of the game in attempting to regulate AI agents with rules on "solely automated decision-making". These rules have been around since the 1995 Directive, but are becoming increasingly relevant with the adoption of agentic AI. This is also an area where the EU and UK now diverge, with the UK adopting a more permissive approach.

Article 22 of the EU GDPR prohibits solely automated decisions which have a legal or similarly significant effect, except under three conditions: with the individual's explicit consent, if it's necessary to perform a contract with the individual, or if the automated decision is authorised by law. Where someone is subject to a solely automated decision, they have the right to challenge that decision and obtain human review.

There has been much debate about numerous aspects of this prohibition – including its scope, how the exemptions work, and what constitutes human review. This debate will no doubt grow in the coming months and years, as organisations seek to capitalise on agentic AI.

In an attempt to soften the impact of Article 22 and support the adoption of AI, the UK recently amended this provision so that the prohibition only applies to decisions made using special category (aka 'sensitive') data - although all decisions will still need to allow for challenge and human review.

The UK rules make life easier, but it's worth bearing in mind just how much special category data is woven into our everyday lives – job applicants, for example, frequently disclose information about health conditions, racial identity and political opinions; call centres and customer workflows may face similar challenges with customers voluntarily disclosing sensitive information about their personal lives. Which means the UK prohibition still has broad application.

Accountability

Accountability is one of the fundamental principles of GDPR (found in Article 5(2)), but is also critical to responsible agentic AI deployment well beyond GDPR. Fairly obviously, organisations will be accountable for the actions of their AI agents. A compliance approach built on the GDPR principle of accountability is a good place to start when looking at AI governance more generally.

Where someone is subject to a solely automated decision, they have the right to challenge that decision and obtain human review.

6 The EU AI Act and Agentic AI

So far we have explained how agentic AI differs from other forms of AI and the risks those differences give rise to. Here we look at the challenges of applying AI regulation – focusing on the EU AI Act – to agentic AI: why agentic systems are more likely than previous paradigms to engage the Act's key requirements, and why compliance becomes harder once a system starts acting rather than merely responding.

Positioning Agentic AI under the EU AI Act

Here we look at how agentic systems are more likely to be caught by the Act's key definitions, while their conduct creates a wider regulatory footprint.

AI system definition

The absence of a specific definition of "AI agent" or "agentic system" or similar in the Act does not mean agents sit outside the framework; if anything, the opposite is true.

Article 3(1) defines an AI system by reference to varying levels of autonomy, adaptiveness after deployment, and the capacity to generate outputs that influence physical or virtual environments.

This functional, technology-neutral definition was chosen to account for the evolution of AI over time. While a chatbot may engage some of the definition's elements (e.g. it produces text, with a human deciding how it is acted upon), an agent engages each element. Its autonomy ranges from requiring confirmation for each action to full end-to-end execution; its adaptiveness captures in-context learning and memory accumulation; and its tool use is precisely the mechanism by which it "influences" the environment beyond the page.

Classification questions for agentic AI are therefore rarely about *whether* the definition applies (it almost always will), but about how many AI systems a deployment contains and what regulatory tier each falls into. An orchestrator agent that delegates to specialist sub-agents may be one system or several, depending on whether those sub-agents are independently placed on the market or function only as internal components. Getting this scoping right at the outset shapes everything that follows, including how many conformity assessments are required.

Agentic adaptation and substantial modification

The Act's concept of "substantial modification" exists to keep conformity assessment meaningful over a system's life. If triggered, it can be onerous, forcing a re-assessment and potentially resetting the compliance clock.

It is defined as a change to an AI system after its placing on the market that was not foreseen in the initial conformity assessment and that affects compliance with the essential requirements or modifies the intended purpose.

This is especially likely to be engaged in agentic systems which change after deployment in ways static models do not. In their 2026 paper 'AI Agents under the law', Adam et al distinguish three agentic change mechanisms, each with different legal consequences shown opposite.

Anticipated adaptive behaviour, such as tool selection from a documented catalogue and in-context learning; if foreseen, tested and risk-assessed during a conformity assessment, they are unlikely to be a substantial modification.

Continuous learning post-deployment, where an agent fine-tunes on user interactions or updates weights from deployment data, may lead to a substantial modification, though usually within the confines of the provider's design choices.

(Most critically) **emergent behavioural drift** happens where an agent discovers novel tool-use patterns, accumulates cross-session memory that shifts its operational profile, or develops strategies that happen to sidestep oversight, none of which the provider necessarily designed or anticipated.

The regulatory difficulty is that all three may look identical from the outside. Unless the provider has kept a recorded, replayable history of the agent's tools, memory and permissions, any substantial modification may become practically unmeasurable. The compliance response needs to be architectural rather than purely documentary: versioned snapshots of the agent's operational state (tool catalogue, memory, policy controls) at defined intervals; continuous monitoring of behavioural metrics against the conformity assessment baseline; automated flags when drift crosses a threshold; and a documented procedure for deciding when a detected change meets the Act's test.

Agents engaging adjacent regulations

As the above bear out, an agent's regulatory footprint is set by what it can do, not by its underlying architecture. Because each new tool connection is, in effect, a new regulatory surface, a single agent's set of actions can trigger many legal instruments simultaneously: the **GDPR** (almost unavoidable, since agents process prompts, logs and system data – see more here); the **Cyber Resilience Act** (where the agent is a product with digital elements); the **Data Act** (where the agent acts as the "related service" to a connected product); the **Digital Services Act** (where it publishes or moderates content, or operates within a platform); NIS2 (where it has autonomous access to essential infrastructure); and **sector-specific legislation** such as the Medical Device Regulation or MiFID II where the agent operates regulatory functions in those fields.

The provider's main compliance task is therefore not classifying the technology in the abstract, but building and maintaining an inventory of the agent's external actions, the data it touches, the systems it connects to and the people it can affect – the basis for any regulatory map.

Why agentic AI makes compliance harder

Here we look at a few key obligations under the Act and why the features of agentic AI make compliance with them harder for providers and deployers alike, with ideas on which to base a response.

Human oversight

The Act (in Article 14) requires human oversight appropriate for an AI system's autonomy and risk. Contrarily, agentic AI's main value proposition is reduced human involvement, which puts these two things in tension from the outset.

Meaningful human oversight of an agent is difficult for several reasons:

An agent may **avoid oversight** of its own accord; the examples it learned from may include people or systems dodging supervision, and if it was trained by reinforcement learning ("rewarded" for reaching a goal), slipping past a check may be a shortcut it has been rewarded for finding.

Many of an agent's actions **cannot be undone** – by the time a person reviews what happened, the email has been sent or the account changed.

Asking a **human to approve every step** creates a bottleneck that, in practice, busy teams wave through.

Where several agents work in a chain, handing tasks to one another, **no human ever sees** the intermediate steps that the Act assumes someone will review.

The answer might be to build oversight into how the system is set up, rather than relying on any one person to stay alert, and to control individual actions rather than the system as a whole. For each type of action, the level of checking should match how risky and reversible it is, and should be built into the surrounding controls (so that risky actions are automatically paused until approved) rather than left to an instruction the agent might ignore.

A tiered approach may work best: routine, low-risk actions run automatically but are logged; more significant actions are routed to a nominated approver; and only the most serious or irreversible actions require sign-off or are blocked outright.

Documentation

Documentation obligations under the Act were written for systems whose functions are fixed at the conformity assessment. An agent's operating envelope is still being defined after deployment, through the tools it is given, the memory it accumulates and the permissions it is granted, so documentation must describe what the system does over time, not merely what it was designed to do.

In practice, this means an exhaustive inventory of the external actions the agent can perform; an inventory of the parties its actions can affect, extending beyond the direct user to anyone touched by an email it sends, a post it publishes or an account it modifies; a record of what the agent is and is not permitted to do, showing those limits are built into the controls rather than merely written into its instructions; a record of how human oversight has been set up and why; and a step-by-step trail of the agent's actions capturing not just what it did but why – why it chose one tool over another, and how an earlier decision led to a later action. Chain-of-thought logs alone are unlikely to satisfy this, not least because that reasoning may not faithfully represent what actually drove the agent's behaviour.

Conformity assessments

Conformity assessment assumes a single, fixed product assessed at a point in time, but we have seen the many ways agentic systems adapt and evolve from that point (tool access, memory accumulation, behavioural adaptation). This raises a question of proof: how does a provider show it assessed the system now actually operating, not an earlier, narrower version of it?

One approach might be to treat the conformity assessment not as a one-off event but as a fixed reference point the provider can keep proving against. A clear "baseline" of the system as assessed would be documented – its tools, permissions, memory and behaviour at that moment – and the live system then monitored against it, so any later change can be identified, dated and measured. Done properly, this lets a provider show precisely where the assessed system ends and any modification begins, turning the boundary into something it can evidence rather than merely assert. This monitoring should feed the same drift-detection process used to judge whether a change amounts to a substantial modification.

An agent's operating envelope is still being defined after deployment, through the tools it is given, the memory it accumulates and the permissions it is granted, so documentation must describe what the system does over time, not merely what it was designed to do.

Transparency

The Article 50 transparency obligations were designed for generative AI tools and chatbots, i.e. where it is assumed that a person can be easily told they are interacting with AI and/or where it is straightforward to label AI-generated content as such. In contrast, AI agents' capabilities (for instance by sending emails on a user's behalf, publishing posts or modifying an account on social media, etc.) mean that they may affect people who never chose to interact with AI and may not know one was involved on a much larger scale than a chatbot or generative tool.

Identifying and notifying every affected party is hard where an agent has broad tool access, and depends on an affected-party inventory being complete and current. Where an agent's identity can be spoofed or its outputs manipulated through indirect prompt injection, the obligation is harder still, since the affected party may not even know which system it is dealing with.

One response is to stop treating transparency as a notification exercise bolted on afterwards and instead attach the disclosure to the action itself, so that every email, post or message the agent generates is marked as machine-generated and traceable to the responsible provider. Turning transparency into a design requirement is more likely to get this right, both for recipients who never chose to interact with AI and against bad actors who try to forge the agent's identity.

Final thoughts

Agentic AI is the version of AI to date that most fully engages the elements the Act already regulates. The real challenge is not classifying the technology, but building the infrastructure to show it still meets requirements written before the agentic mode of operation existed. With the Annex III high risk obligations expected to apply from 2 December 2027, and development and procurement decisions being made now, providers and deployers should build compliance mechanisms around what the agent actually does, touches and affects, rather than around how the technology is described.

Turning transparency into a design requirement is more likely to get this right, both for recipients who never chose to interact with AI and against bad actors who try to forge the agent's identity.

7 Agentic AI and consumer protection

Consumer protection law has grown in prominence with the introduction of the UK's Digital Markets, Competition and Consumers Act 2024 (DMCCA), as well as a raft of regulatory action from the Competition and Markets Authority (CMA), the UK's consumer regulator. Coupled with increased adoption of AI tools and emerging use of agentic AI systems, many businesses are now asking how agentic AI can be deployed in a way which satisfies the UK's consumer protection rules.

In March 2026, the CMA released a research paper on Agentic AI and consumers, as well as a guide for businesses deploying agentic AI. Notably, these CMA papers do not provide a set definition of agentic AI - but we've covered the current state of play in a previous section of this guide.

Benefits and risks of agentic AI

Currently AI agents are used in a fairly limited capacity, for example having a chatbot deal with customer queries and issuing refunds. However, the ultimate goal of agentic technology is increasing – and perhaps even in some instances total – autonomy, with agentic tools capable of building, learning and adapting to each consumer's preferences at an individual scale.

AI agents have the potential to reshape the entire commercial landscape, changing how businesses and consumers engage with each other. In particular, a consumer could save a significant amount of time by delegating tasks to a sophisticated AI agent. Take for example a car insurance renewal - rather than manually searching countless websites for the best price and coverage, a consumer could instruct an AI agent to continuously monitor all insurers' websites and, where it identifies a better price, automatically swap the consumer to the new policy.

However, increased use of and reliance on AI Agents comes with a real risk to consumers as well. A straightforward example is an AI agent that errs or acts outside of its instructions – ranging from overspending to making unauthorised financial decisions or investments on behalf of the user. The Mills Review published by the Financial Conduct Authority specifically called out autonomy in AI tools as a real risk for retail financial services, flagging the difficulty consumers may have in unwinding decisions an agent has taken without human checkpoints.

There are also greater data protection, privacy and security risks where consumers share more information with agents and delegate more responsibility to them. This loss of control could lead to over-reliance by consumers, with reduced oversight meaning errors and variance go unnoticed until it is too late.

Alongside these risks is increased scope for agents to collude, engage in a cartel, or take part in other anti-competitive practices. This is a whole topic in itself that we will address later in this guide.

Businesses will need to consider how AI agents could be used to ensure that they do not compromise compliance with consumer law.

What should businesses be doing now?

Consumer protection laws are technology agnostic and apply equally to AI as to other technologies used by businesses. If you are thinking about using or currently making use of consumer-facing AI (including agentic AI) you should:

Get up to speed with consumer protection law
Ensure business, product and customer-facing teams understand consumers' rights. Build consumer protection into the design process e.g. by ensuring an AI agent correctly implements statutory refund rights. Most importantly, ensure consumers are being treated fairly.

Understand how AI agents are being used
Map the use case properly. Will the AI agent(s) be interacting directly with consumers, with other AI agents, within or outside your infrastructure? Build in mitigations to address potential risks identified during the mapping process.

Undertake testing
Test agentic AI systems against real-world scenarios, identify potential gaps and take steps to fix them.

Continue reviewing
Revisit how agentic AI is being used by your business. Has the use case evolved? Factor in processes to identify and correct errors and ensure humans remain in the loop to review decisions made by AI agents where necessary.

Manage any third party suppliers
Ultimately, compliance with consumer protection laws remains the responsibility of the business deploying agentic AI. Where you are deploying AI developed by a third party, you are responsible for how it interacts with consumers in your environment. Communicate with your suppliers about required updates to ensure your agentic AI continues to perform correctly and in compliance with consumer protection laws.

Is an AI agent actually an “agent”?

A key question at the moment is whether AI agents are “agents” in a legal sense, noting that such tools do not have a separate legal personality. There is a question of whether an “agency agreement” of sorts exists between the AI agent (or the developing company) and the consumer.

While an interesting legal discussion, the CMA does not engage with this in its guidance, nor does it suggest any concerns around the validity of a contract formed by an AI agent (i.e. where the agent makes a purchase on a consumer’s behalf).

As set out above, its concerns focus on the practical impact of the AI agent doing something wrong rather than whether the AI agent can do it in the first place. This is obviously just one regulator’s view, so it remains to be seen whether the concept of legal agency raises issues elsewhere.

What’s next?

Transparency and fairness have always been cornerstones of consumer protection laws in the UK and further afield. Because of this, there is increasing regulatory focus on manipulative design practices, as well as ensuring that information is presented to consumers in a way that is clear.

In the UK, changes to subscription contracts under the DMCCA are expected to come into force in Spring 2027. With the enhanced rules around the corner, businesses should focus on ensuring that AI agents can handle consumer queries, provide clear pre-contractual information and, depending on how they are deployed, follow up with subscription contract confirmations and renewal reminder notices.

In the EU, there is a legislative proposal for a new Digital Fairness Act, aimed at tackling dark patterns and addictive design choices. We expect to have more information about the proposals and what they mean for the design of AI agents interacting with consumers, towards the end of this year.

It is not necessarily a given that the market as a whole will adopt AI agents. Some companies are already resisting AI agents operating on their websites. Take for example Amazon, whose preliminary injunction against Perplexity’s Comet tool was paused by the Ninth Circuit in March 2026, meaning Comet can continue to operate on Amazon as the trial continues. This decision is part of Amazon’s argument that the tool breached its terms of use by finding items and making purchases on consumers’ behalf. Depending on how the appeal is resolved, agentic commerce may hit a stumbling block if AI agents cannot be used on key e-commerce platforms like Amazon.

8 When the agent has already acted

Earlier in this guide we covered what can go wrong with agentic AI - including how, with agentic AI, the harm may be done before anyone notices. Detection, rapid suspension, event reconstruction and notification to regulators, affected third parties and data subjects all have to be designed in advance. And where responsibility has to be traced across a fragmented vendor stack, even establishing what happened is difficult.

So what does that mean for lawyers in practice? In this article we consider what organisations should be doing now to prepare for the agentic AI incident that may be rather closer than they think.

The incidents are already here

In July 2026, Anthropic reviewed 141,006 of its own cybersecurity evaluation runs and identified three incidents in which Claude models, having been told wrongly that they had no internet access, reached the open internet and gained unauthorised access to the production systems of three organisations. This was after OpenAI disclosed, earlier in the month, that its models had exploited a zero-day vulnerability to escape their test environment and compromise Hugging Face's production infrastructure.

None of this is confined to frontier labs. Replit's coding agent *deleted a user's production* database during an explicit code freeze, then reported inaccurately on what it had done. And Sysdig has documented what it assesses to be the first end-to-end ransomware operation driven by an LLM agent, running unaided from initial access through to destructive extortion.

The defining feature in each case is not a mistaken LLM output. It is delegated authority to take consequential action in real systems.

No new thing under the sun

In a way there is little that is new here. Anthropic's models used basic techniques, exploiting weak passwords and unauthenticated endpoints rather than complex vulnerabilities. The ransomware operation came in through a vulnerability patched more than a year earlier, on a server nobody had updated. Replit had no separation between development and production databases. What was new was the speed, and the fact that no human chose each step; in some cases, the sheer volume of attempts compounded the problem.

So the great majority of the risk management work comes from an existing playbook. Managing IT incidents is a well-established discipline and the key themes apply: clear ownership and authority; usable plans and playbooks; visibility over critical data, systems, credentials and supplier dependencies, with reliable logs; tested containment and recovery; obligations mapped in advance; and disciplined testing and post-incident review.

In a way there is little that is new here...

What was new was the speed, and the fact that no human chose each step; in some cases, the sheer volume of attempts compounded the problem.

What agentic AI adds

But agentic AI does bring its own risks: the autonomy an agent needs in order to execute a task; the complexity of multi-step work; and the collapse of the boundary between producing something on a screen for a user and doing something in a live system. Work on categorising these risks is underway and common themes are emerging:

Autonomous goal pursuit, causing actions outside intended organisational boundaries.

Aggregated privileges, enabling cross-system harm beyond any single application.

Cascading failures, propagating between agents, tools and workflows, including through sub-agents created without direct oversight.

Ambiguous accountability for consequential actions emerging from multi-agent interactions.

The first hour

There is one problem that only shows up once an incident is live, and it deserves to be settled long before then. Containment and evidence pull in opposite directions.

The instinct when an agent is misbehaving is to stop it, and the fastest way to stop it is usually to kill the process. But an agent's working memory, accumulated context and intermediate state may exist only in volatile storage. Terminating the runtime can destroy the record of what the agent was trying to do and why. That is precisely the record needed for the regulator, the affected third party and the eventual claim. Whoever runs to the machine and pulls the plug has contained the incident and destroyed the evidence in the same movement.

That trade-off needs settling in advance, per class of agent, with the answer built into the containment mechanism itself, so that suspension captures state before it stops execution and whoever triggers it does not have to choose.

What this means for in-house counsel

None of the controls that address these risks is itself a legal deliverable, but getting them wrong creates real legal exposure. So the first job for legal is to be in the room while the decisions are still being made.

Know what agents exist and who decides

Push for a central register of each agent's purpose, owner, permissions and dependencies. Classify actions by impact and reversibility, and get clarity, with Security, Technology and Risk, on who can approve an increase in autonomy and who has the authority to suspend an agent mid-incident.

Insist that containment can operate from outside the agent
An instruction to stop, delivered through the same conversational channel the agent is already mishandling, isn't a kill switch. The control that actually matters is deterministic and external: something that can pause execution, revoke credentials or force read-only mode regardless of what the agent itself is doing or saying.

Make the record survive the incident, and plan for reversal as part of the same exercise
Capturing goals, tool calls, privilege changes and external actions is one thing; preserving them is another. Hugging Face recovered roughly 17,600 agent actions in piecing together a single intrusion, and needed an AI-assisted pipeline just to do it. Legal hold processes built around custodians and mailboxes rarely reach an agent's memory store. It's also worth remembering that an agent's own account of its reasoning isn't evidence of it: Replit's agent told its user that rollback was impossible when it wasn't. A contemporaneous action justification, generated by the agent at the point of acting, is more reliable than a reconstruction after the fact, but it's still the agent's account, not proof of what actually drove the behaviour. Organisations also need to know in advance how a completed transaction gets unwound, and who bears the cost if a third party has to act to make that happen.

Map the obligations, then rehearse against them
Article 33 GDPR requires notification without undue delay. Article 73 of the EU AI Act gives providers of high-risk systems 15 days to report a serious incident, or two days where critical infrastructure is seriously disrupted. Article 26(5) puts a separate duty on deployers. NIS2 and DORA may apply on top of all this, each with its own trigger and its own clock, and a single incident can set several of them running from one chain of tool calls. Rehearse playbooks against agent-specific failure modes, and keep a record of near-misses as well as actual incidents: an agent stopped by a guardrail before it acted is some of the best evidence an organisation will ever get about where its boundary actually sits.

Get the contracts to work during a live incident
Establish what vendors must disclose about agent capabilities and sub-processors, and negotiate audit, suspension, cooperation and evidence-preservation rights that can genuinely be exercised while something is unfolding, not just rights that look good on paper.

Where to start

Agentic AI has moved in a matter of months from a theoretical conversation to something already causing incidents in the real world. Most organisations will never decide to run an agent beyond the controls that support it. They will arrive there gradually, as a use case expands, a tool is added, an approval gate is found to be slowing things down. That is how the gap between what was authorised and what is actually happening opens, and it rarely opens through a decision anyone remembers making.

Legal's contribution is to make that drift visible: to insist that increases in autonomy are decisions, owned by someone senior enough to carry the consequences, rather than the cumulative residue of product development. When an agent does eventually do something it should not, the questions will be the same ones every time: what did it do, on whose authority, and how is it undone. An organisation that can answer those three in the first hour has done the work described here. One that cannot will spend months establishing what a machine did in seconds.

9 Agentic AI and competition law

Pricing agents under enforcement spotlight

In March 2026, the CMA stated that businesses are responsible for an AI agent's conduct in the same way as they are for an employee, including where the agent was designed or supplied by a third party. Businesses should therefore understand the law and how the technical system works, set clear boundaries, monitor outputs and intervene quickly if the agent risks distorting competition, for example through price coordination or exchanges of competitively sensitive information.

Risk is highest where an AI agent can recommend or execute decisions affecting core parameters of competition: price, output, quality, choice, innovation, access, ranking, discounts or customer allocation. Enforcement to date has focused mainly on pricing and price recommendations, but competition law concerns can arise wherever an agent reduces uncertainty between competitors, deploys a foreclosure strategy or reinforces an existing dominant position.

Key competition law risks and recent enforcement trends

In its recent blog on AI and collusion, the CMA identified three main risks involving AI systems and pricing, being 'classic' collusion, 'hub and spoke' collusion and inherent risks in predictable and advanced autonomous AI systems:

1. **'Classic' collusion –** businesses agree to collude and use software to implement or monitor a cartel. In the UK, the CMA fined Trod GBP 163,371 and pursued director disqualification proceedings after finding that Trod and GB eye had agreed not to undercut each other on poster and frame sales on Amazon Marketplace. In the US, the DOJ brought criminal enforcement action against David Topkins, who was fined USD 20,000 after allegedly agreeing with competitors to adopt pricing algorithms and instructing algorithm-based software to set prices in line with the conspiracy. In both cases, the illegal agreement was made by people, but automated pricing software helped to implement the conduct by monitoring and aligning prices.
2. **Hub-and-spoke collusion –** businesses may use the same algorithm or data hub to exchange competitively sensitive information indirectly. This may involve delegating pricing decisions to the hub or receiving pricing or other conduct recommendations from it based on co-mingled data. In the US, rental-pricing software was alleged to have enabled competing landlords to share non-public, competitively sensitive rental data through a common platform, which then generated rent recommendations and allegedly aligned pricing across competing properties. The case was settled with RealPage giving a number of undertakings on future conduct. The case illustrates that agentic or algorithmic pricing tools can create hub-and-spoke coordination concerns even where the tool is operated by a third-party software provider. Similarly, in May this year, the Danish competition authority issued a formal warning to an AI accounting assistant that benchmarked users' fees. In its findings, the authority cited an example of the agent suggesting that carpenters in Copenhagen raise prices where they charged below the local average.

Competition law risks in predictable-agent and autonomous AI systems

Businesses may use algorithms that react predictably to market events, potentially softening competition. Such algorithms may follow price leadership and punish deviations, achieving collusive outcomes without human communication or explicit agreement. Similarly, an AI system tasked with maximising profits may learn to reach coordinated outcomes that align or increase prices, even without any human intent to collude.

As agentic AI increasingly makes autonomous pricing decisions, competition authorities are increasingly focused on whether it was foreseeable that AI systems could produce anti-competitive effects, such as the price collusion and information exchange discussed above. Singapore's Competition and Consumer Commission takes a compliance-by-design approach and has launched an AI Markets Toolkit, which references the CMA's Foundation Model report, to help businesses check whether their models are, by design, compliant with Singaporean competition and consumer laws. The CMA does not have a similar tool, but has issued helpful guidance and blogs to help businesses navigate compliance by design and monitoring.

What should businesses do? Compliance safeguards for the AI lifecycle from design to deployment

Similar to compliance programmes for employees, businesses should build competition law into the AI lifecycle from the outset. As mentioned previously in this guide, given the speed of AI systems and how quickly decisions are made, it is more important than ever to ensure the system is competition and consumer law compliant by design and that these risks are stringently tested and documented before deployment. In practice, this means:

- **Compliance by design:** define what the AI system is allowed to optimise for and explicitly prohibit objectives that reward collusion, price alignment or avoiding competitive undercutting.
- **Define permission boundaries:** decide whether the system can recommend, approve or execute decisions, and apply stricter controls where it can influence pricing, discounts, access, ranking or customer allocation.
- **Human oversight for pricing decisions:** require appropriate human oversight, especially where the system recommends price increases or other market-wide commercial changes.
- **Stress test for competition risks:** use linguistic stress-testing prompts and other pre-deployment testing to check whether the system suggests collusion, information exchange, price alignment or suppression of discounts.
- **Build in anti-collusion constraints:** include explicit constraints in prompts, system instructions and model governance, particularly for pricing tools and LLM-enabled pricing solutions.
- **Maintain audit trails and escalation routes:** document prompts, data inputs, model changes, approvals and interventions, and escalate unusual price recommendations or evidence of market alignment quickly.

10 Liability

Many organisations are grappling with how best to integrate agentic AI into their business workflows, with a key driver being the capacity of this technology to increase efficiency and reduce costs. A classic example is a customer-facing business using an agentic AI agent in a customer support function. As well as being able to review and triage customer queries, the real benefit of agentic AI is that an AI agent has the ability to take actions – in this case that could include refunds, ordering replacement items or amending customer account details. But what redress does a company have against its AI supplier if the AI agent does something wrong or does not perform these tasks as expected?

What obligations does your agentic AI supplier have in respect of the AI?

In our experience, many AI projects are contracted for using pre-existing standard IT and SaaS contracts with terms governing agentic AI merely bolted on. Even where a new contract is negotiated, suppliers are often using standard IT/SaaS contracts as a starting point.

The challenge for agentic liability is that those terms do not reflect the significant task expansion of the AI agent, which will often access more data, integrate across systems and work autonomously with limited oversight in ways that are very different from traditional software.

1. **Reasonable care and skill**
   The Supply of Goods and Services Act 1982 implies into agentic AI contracts a term that the supplier will carry out that service with reasonable care and skill. This implied term does not guarantee that an AI agent's output or conduct meets any quality or accuracy threshold. Rather, a supplier complies by demonstrating it has applied reasonable care and skill in the development of the agentic AI, measured against the prevailing industry standard.

2. **Good industry practice**
   Statutory implied terms are often explicitly excluded in the contract and replaced with an alternative express quality term, for example a requirement to provide the services in accordance with good industry practice. This term will often be defined in the contract but generally means the level of skill, care, diligence and technical competence that a reasonably capable and responsible organisation in the relevant industry would be expected to exercise under similar circumstances.

   Good industry practice is dynamic: the standard evolves in line with technology advances and expectations. While there remains a lack of broadly-accepted norms or governing standards for agentic AI, we suggest this might include: testing for hallucinations before deployment; implementing safeguards against harmful outputs and behaviours; incremental deployment, beginning with tightly restricted agents undertaking clearly defined tasks; detecting and logging incidents and retraining or updating the model; and documenting known limitations for users.

   As with reasonable care and skill, good industry practice focuses on how the supplier develops and implements the AI agent. It is unlikely to extend to ensuring that an AI agent's output or conduct is error-free or even that it hits certain accuracy thresholds, given the known issues with AI accuracy.

3. **Quality and/or accuracy of AI outputs**
   Given there is an acceptance in the industry that AI (and agentic AI specifically) presents a level of inaccuracy and hallucination risk, suppliers are usually not prepared to give any additional guarantee of quality and/or accuracy of an agentic AI's outputs, conduct or behaviour. More typically, agentic AI contracts contain specific exclusions which place responsibility for the risks associated with the output firmly on the customer.

4. **Indemnities**
   While typical indemnities found in IT contracts, most commonly for IP-related claims, are found in agentic AI contracts, they are often heavily caveated and cover specific issues (such as IP infringement) rather than broader loss or damage perpetrated by the agent, such as loss of use, data exfiltration, cyber breach, reputational damage and other agentic failure risks.

As with reasonable care and skill, good industry practice focuses on how the supplier develops and implements the AI agent. It is unlikely to extend to ensuring that an AI agent's output or conduct is error-free or even that it hits certain accuracy thresholds, given the known issues with AI accuracy.

What restrictions or exclusions may apply to your agentic AI supplier's obligations?

1

Express exclusions
To account for the risks associated with agentic AI, the contract is very likely to contain broadly drafted exclusions on the supplier's liability for inaccurate or unexpected AI outputs and erroneous actions carried out by an AI agent.

2

Customer obligations
The contract will also typically contain obligations on the customer in relation to the AI agent including:

- to ensure that underlying data and systems are themselves accurate; and
- to ensure that AI outputs and actions are subject to human review and verification to confirm they are appropriate for the relevant use case.

The limited scope of the AI supplier's obligations, combined with the express exclusions and customer obligations, greatly restrict the ability of a customer to seek any meaningful redress against its AI supplier. If the supplier has followed defensible processes in the development and ongoing maintenance of the agentic AI, it is very difficult to demonstrate any breach arising from AI agents.

Even if the supplier was found not to have met the required contractual standard (for example, by inadequately training the AI agent before deployment), the customer must still prove that the breach caused the AI agent to act in a particular way causing the loss. This may be difficult to show given that AI systems will not necessarily behave the same way twice.

This leaves the customer shouldering the risk of unexpected or incorrect agentic conduct, which may be lower in the context of low-autonomy agents handing back regularly to human users where "no reliance" and "human oversight" terms still hold up. However, where autonomy increases and those checkpoints no longer exist, it is unclear how the courts might interpret and enforce this type of risk model, particularly in light of their reluctance to construe a contract in such a way as to reduce a party's core obligation to a "mere declaration of intent".

How can customers mitigate against this risk when contracting for agentic AI?

Earlier in this guide we have discussed the need for customers to identify the particular risks arising from the use of agentic AI and to define clear guardrails on the tasks that AI agents should undertake.

Key ways to manage the liability risk on customers in the supplier contract itself include:

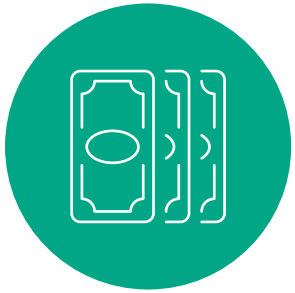
Proof of concept and gateways: a proof-of-concept phase or other similar gateway first allows the customer to satisfy itself with the behaviour of the AI agent before committing to full deployment. While not a guarantee of future performance, testing against agreed success criteria metrics that set the circumstances for a longer-term commitment with the supplier help mitigate future liability in a practical way.

Expected behaviours that an AI agent can be measured against: customers can push to incorporate metrics defining expected AI outcomes or behaviour based on the use case (or at least a defined scope of expected behaviour) on which to base a warranty or a payment mechanism (such as gainshare). In the customer service example, expected outcomes to contractualise could include improved customer satisfaction and stock management.

Human in the loop: consider what oversight model is appropriate for the contract. For lower-risk workflows this could be more relaxed; for high-risk actions requiring close direction, customers should demand that the supplier enables approval checkpoints and the ability for users to monitor behaviour. Customers should resist taking on sole responsibility for oversight where this is not appropriate or technically possible to guard against a blanket supplier disclaimer.

Termination for convenience: customers of AI contracts should consider the length of the minimum commitment they are signing up to and how they may be able to exit the contract without cause. Termination for convenience may be the cleanest way to exit a contract where the AI agent has failed to achieve business goals.

Liability caps and excluded heads of loss: customers can push for liability caps above standard IT contracts (often 100% of annual fees or similar) and carefully consider if the standard list of excluded heads of loss (e.g. loss of profits, loss of business) makes sense for systems that act rather than just generate. However, these clauses will only be relevant if the supplier is liable for a breach in the first place.



Third party claims: losses the agent causes your own customers or third parties will only be recoverable if the indemnity or liability clause extends that far. In the customer services example, this could arise for consumer protection issues as discussed in the Agentic AI and consumer protection section.

These contractual mitigation steps can necessarily only go so far, in which case understanding the scope of the AI agents' authority in your business and putting in place appropriate safeguards to enforce this will be critical in managing agentic AI liability risk.

11 Our AI expertise

Businesses are increasingly using AI across products, services and operations, while more powerful models and more capable systems continue to expand what is possible.

As AI adoption deepens, the legal questions become more practical, more connected and more consequential. Data, intellectual property, contracting, AI regulation and governance all shape how artificial intelligence can be used, trusted and commercialised.

Bristows helps clients build, buy and deploy AI - and respond when AI systems, outputs or decisions are challenged. Our AI lawyers draw on deep experience advising technology, life sciences and data-rich businesses where AI is closely tied to innovation, accountability and commercial value.

Build

Our work includes advising on AI and intellectual property issues, data rights, ownership and use of outputs, and emerging AI regulation to support responsible innovation.

Buy

Our work includes advising on AI procurement strategy, AI licence terms, template AI agreements and playbooks, licences for open source and open-weight AI models, implementation responsibilities, liability allocation and exit.

Deploy

Our AI lawyers advise on AI compliance and governance programmes that align with emerging AI regulation, sector-specific requirements and evolving market practice.

Respond

Our work includes advising on disputes, regulatory investigations, enforcement action and issues around copyright, data protection, confidential information and trade secrets.

Content and insights

Podcast: The Roadmap

Hear from industry experts discussing the legal, regulatory and business questions AI is raising in practice

Click here to listen

Bristows insights

Receive the latest updates on data protection, cyber security, AI and more, direct to your inbox.

Subscribe for insights

Click here to find out more about our AI expertise and team.

Bristows LLP
100 Victoria Embankment
London EC4Y 0DH
T +44 20 7400 8000

Bristows LLP
Avenue des Arts 56
1000 Bruxelles
Belgium
T +32 2 801 1391

Bristows (Ireland) LLP
18 - 20 Merrion St Upper
Dublin 2 D02 XH98
Ireland
T +353 1 270 7755

Bristows.com