



From model risk to AI risk

Building trust in intelligent systems

forv/s
mazars

Contents

3	Executive summary
4	AI regulatory landscape AI model risk management shortcomings
6	Identification and risk assessment Agentic systems and the model definition The agentic system as the risk object
9	Governance
10	Development Fairness Explainability Frontier model configuration
12	Validation Model drift and revalidation Agentic system validation Third-party and vendor AI models
16	Risk mitigation framework for AI Tools for risk mitigation
18	How can Forvis Mazars help?

Executive summary

Artificial Intelligence (AI) is reshaping how organisations operate – enhancing decision-making, automating processes, and unlocking new sources of value. From customer engagement to risk assessment, AI tools are increasingly embedded in core business activities and strategic initiatives.

This report explores model risk management in relation to models which use AI techniques to produce outputs that influence business decisions, processes, customer interactions, financial reporting or risk management activities. In contrast, this article does not cover AI tools which are being increasingly used to improve efficiency, i.e. Copilot, but may also require proportionate governance, oversight and control. Increasingly, however, AI is deployed not as a single call to a model but as an agentic system, in which a frontier model plans, selects tools and takes actions with a degree of autonomy. Such systems require their own classification treatment, which this article addresses directly.

As adoption accelerates, so too does the complexity and scale of associated risks. AI models are exposed to many of the challenges associated with traditional models, including data quality issues, model drift and governance weaknesses. However, the scale, complexity and uncertainty associated with AI can amplify these risks, while also introducing new ones.

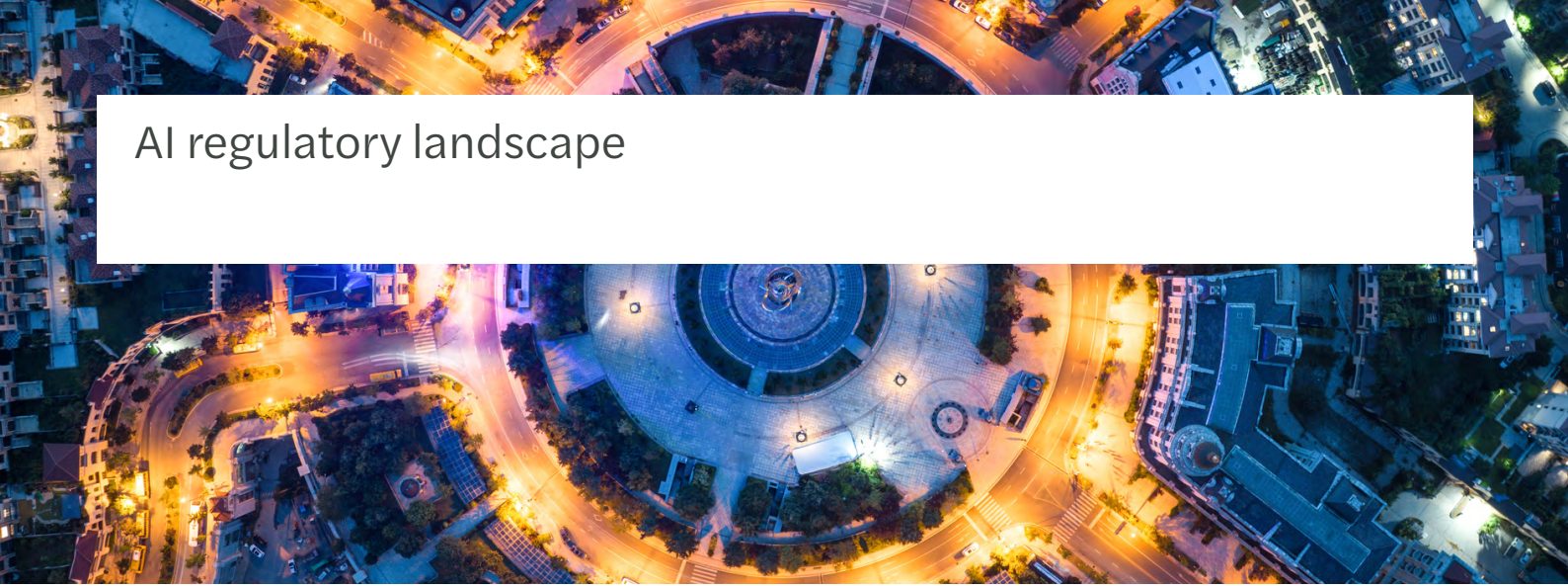
Key challenges faced by AI systems include:

- Model drift and continuous learning, requiring ongoing oversight rather than one-time validation.
- Limited transparency and explainability, particularly for complex AI models, making outcomes harder to interpret and justify.
- Data dependency and bias risks, potentially leading to unfair, discriminatory or unintended outcomes.
- Evolving regulatory expectations, with increasing scrutiny on governance, accountability, and ethical use of AI.

The enhanced traditional risks, combined with the AI specific risks introduced, mean that existing risk management approaches are no longer sufficient on their own. Organisations must evolve their frameworks to address the distinct characteristics of AI while remaining consistent with established risk management principles.

Most banks and insurers are generally not creating standalone AI risk frameworks. Instead, they are integrating AI into existing risk management disciplines such as model risk management (MRM), operational risk, conduct risk, data governance, cyber risk, operational resilience and Senior Managers and Certification Regime (SM&CR) accountability. This is consistent with the stated approach of the Financial Conduct Authority (FCA) and Prudential Regulation Authority (PRA), which continue to rely primarily on existing frameworks rather than introducing a comprehensive AI-specific regime.

AI is scaling rapidly – but so are the risks.
Organisations must evolve risk management to ensure AI is trusted, controlled, and compliant.



AI regulatory landscape

The global regulatory landscape for AI continues to evolve. While the EU has introduced a dedicated regime through the [EU AI Act](#), the UK has adopted a principles-based approach that relies on existing regulatory frameworks rather than AI-specific rules.

For UK financial services firms, the FCA, PRA and Bank of England expect AI-related risks to be managed through existing governance and control frameworks. Areas of focus include model risk management, data quality, operational resilience, third-party dependencies and accountability.

Although a standalone UK financial services AI regime has not been introduced, firms should expect increasing regulatory scrutiny and be prepared to demonstrate that AI risks are identified, assessed and managed in a proportionate and effective manner.

For financial services, the [PRA's SS1/23](#) on model risk management provides a key reference point for AI models that fall within the scope of a firm's model inventory. For insurers and other firms, the principles may still be relevant as good practice, particularly where AI is used in material decision-making, pricing, risk assessment or operational processes.

This report introduces a structured framework designed to:

- Provide accountability across the AI model lifecycle.
- Enable consistent identification, assessment, and mitigation of AI model risks.
- Support proportionate controls and independent validation.
- Ensure alignment with emerging regulatory and ethical standards.

In line with PRA and FCA expectations, this framework considers existing regulation and enables an organisation to scale AI confidently and responsibly, balancing innovation with control to deliver trusted, transparent, and compliant AI-driven outcomes.

AI regulatory landscape

AI model risk management shortcomings

In October 2025, the PRA held two CRO roundtable sessions to discuss the adoption of AI and Machine Learning technologies. Below we consider the recurring gaps the supervisor observed as firms adopt AI under [SS1/23](#):

- **Risk appetite set too late - or not at all**
Boards should set a model risk appetite, but AI's opacity means firms deploy before expressing a tolerance, so they can't tell when a model has exceeded it. Setting appetite before deployment is what keeps AI within bounds.
- **Tiering policy and inventory don't match** The PRA observed firms whose policy sets AI/ML at a minimum medium complexity, while the inventory tagged those same models low. Inaccurate inventories give an incomplete view of aggregate model risk.
- **Aggregate risk is underestimated**
A model that looks immaterial in isolation becomes material once the same AI technique is deployed across many portfolios or jurisdictions, yet tiering approaches aren't being challenged for this.
- **A market-wide skills & challenge gap**
Firms lack the blended quant, technology and actuarial capability, across all three lines of defence, to genuinely challenge AI models rather than rubber-stamp them

The shortcomings observed by the PRA highlight a common theme: while AI adoption is accelerating, many organisations' risk management capabilities are not evolving at the same pace. Against this backdrop, understanding the emerging regulatory landscape is critical. Regulators are increasingly focused on how firms govern, monitor and control AI, expecting existing risk management frameworks to be adapted to address AI-specific risks..



Identification and risk assessment

Not all AI tools should be treated as models. AI models are defined as models that use machine learning, generative AI, natural language processing, computer vision or other AI techniques to produce outputs that influence business decisions, processes, customer interactions, financial reporting or risk management activities (Zones 2 & 3 in the diagram below). This contrasts with other AI tools such as Microsoft Co-pilot, Chat GPT etc which are used to improve productivity and efficiency (Zone 1).

Applying a uniform model governance approach to every AI use case leads to unnecessary complexity, increased costs, and operational bottlenecks. Instead, organisations should clearly classify AI functions – distinguishing decision-making models from assistive tools – and apply a tiered, risk-based approach. This ensures that rigorous governance is focused on high-impact, high-risk use cases, while lighter controls are applied where appropriate. The result is faster time-to-value, more efficient controls, and the ability to scale AI safely in line with evolving regulatory expectations.

Effective AI risk management will involve clearly identifying which tools fall within the scope of the PRA's definition of a 'model' and consistently classifying their risk. This ensures governance is proportionate, targeted, and focused on the areas of greatest impact.

To support proportionate and effective risk management, AI models are assessed and categorised into distinct risk zones, reflecting their potential impact on customers, operations, and regulatory exposure. This framework is primarily designed to address the requirements of Zone 3 (high-risk AI) where the potential for harm or regulatory scrutiny is greatest. By establishing a robust approach for Zone 3 models, an organisation can set a consistent benchmark for practice and ensure confidence in the most critical AI-driven decisions. Starting with the highest risk use cases, organisations can scale proportionately across their AI portfolio.



Identification and risk assessment

In line with [SS1/23](#), firms should maintain a comprehensive, firm-wide model inventory to enable the identification of models, sources of model risk and provide the necessary MI for reporting on model risk.

A key challenge is ensuring the model and AI inventory remains complete, current and consistently classified as new use cases emerge. Firms should establish clear ownership, periodic attestations from business areas, and risk-based reviews to identify unregistered models, AI tools or end-user solutions. There are useful lessons from the historical governance of spreadsheets: organisations learned that innovation thrives when users are empowered to develop solutions, provided proportionate controls exist around inventory management, materiality assessments, validation and ongoing monitoring. The same principle applies to AI, where the objective is not to constrain innovation but to enable it safely through appropriate guardrails. Internal audit, independent risk functions, consultants and increasingly AI-enabled monitoring tools can all play a role in performing periodic assurance. For example, automated discovery capabilities can scan code repositories, applications, workflows and technology estates to identify potential AI usage, helping firms detect shadow AI, validate inventory completeness and target reviews towards higher-risk implementations.

Forvis Mazars helps organisations establish and maintain complete AI and model inventories through risk classification frameworks, inventory reviews, and governance processes. We support firms in determining which AI use cases fall within scope of model governance, applying proportionate risk tiering and identifying gaps in existing inventories to reduce the risk of ungoverned deployments.

Agentic systems and the model definition

A frontier model embedded in an agentic system should first go through an MRM classification assessment before its treatment is decided. Whether it formally enters the model inventory depends on how the organisation defines a model and how the system is used, applying the same PRA expectation set out above, that firms use judgement to identify models and apply controls proportionate to materiality.

An agentic system is likely to fall within MRM where it:

- Recommends or makes credit, pricing, underwriting, claims, fraud, capital, compliance or customer decisions.
- Produces scores, classifications, estimates or judgements.
- Influences a material business or regulatory outcome.
- Operates with meaningful autonomy.
- Can change its actions based on context, memory, tools or previous results.
- Is difficult to explain, reproduce or independently validate.

It may not meet the strict model definition where it only summarises documents, retrieves information or performs low-risk administrative tasks. Such systems still require AI governance, operational risk, technology risk, third-party risk, data protection, cyber security and conduct controls, applied proportionately to their actual risk profile rather than waived by default.

Identification and risk assessment

The recommended governance position is that all material agentic systems undergo MRM classification. Systems that meet the model definition enter the model inventory. All others enter an AI system inventory and remain subject to equivalent risk-based governance. This dual-track approach closes a specific and avoidable gap, the argument that a system is 'not technically a model' and therefore sits outside proper oversight. The [US Federal Reserve's SR 11-7](#) guidance takes a similarly broad, risk-based view of models and systems used in decision-making, which supports treating this as sound practice rather than a UK-specific position.

The agentic system as the risk object

The frontier model is only one component of an agentic system. The real risk object is the complete system in which that model operates, including the context it is given, the harness that directs its reasoning, the memory it retains, the tools and APIs it can call and the guardrails that constrain it. Traditional model validation, applied to the frontier model alone, will not be enough.

MRM therefore needs to assess each layer of the system in its own right:

- **Frontier model:** limitations, version changes, bias, hallucination and vendor dependency.
- **Context:** system prompts, retrieval-augmented generation sources, data quality and permissions.
- **Agent harness:** planning, reasoning loops, tool selection and delegation.
- **Memory:** stored state, incorrect memory, privacy and contamination.
- **Tools and APIs:** what the agent can read, change, approve or execute.
- **Guardrails:** policy checks, limits, escalation and human approval.
- **Outcomes:** accuracy, stability, explainability, and customer and financial impact.

Assessing the frontier model in isolation, however rigorously, leaves six of these seven layers unexamined. A firm that validates the model and stops there has not validated the system that customers and the business actually rely on.



Governance

Robust governance is critical to ensuring AI models are developed, deployed, and operated in a controlled, transparent, and accountable manner. A governance framework establishes clear ownership, oversight, and decision-making structures across the AI lifecycle.

The governance of AI models differs from that of a traditional model in a number of ways, including a cross-functional governance requirement, a shift in the governance of outcomes to the impact on individuals and bias, and continuous monitoring due to sustained learning.

AI model governance is underpinned by the following principles:

- **Accountability:** clear ownership for model outcomes and risks.
- **Proportionality:** controls aligned to model risk (based on classification).
- **Transparency:** the ability to explain and justify model behaviour and decisions.
- **Independence:** appropriate separation between model development and validation.
- **Lifecycle oversight:** continuous governance from design through to retirement.

Governance establishes the accountability and control structure for AI models. These principles must be consistently applied throughout each section of the model lifecycle, including the development stage during which key decisions concerning the risk appetite and training data for the model are undertaken.

In practice, firms should map AI oversight into existing governance arrangements, with clear ownership across the first line, second line and independent validation functions. Governance forums should be able to evidence how AI risks are escalated, challenged and monitored against risk appetite.

Forvis Mazars supports organisations in designing and embedding effective AI governance frameworks that balance innovation with appropriate risk oversight. Our experience spans policy and framework development, AI model inventories, risk classification methodologies, governance committee design, escalation processes and independent challenge mechanisms. We also assist in operationalising governance throughout the AI model lifecycle from development through to monitoring, validation and retirement.

Development



Under [SS1/23](#), model development must be supported by:

- A clear statement of purpose and design objectives.
- Justified choice of methodology.
- Evidence of testing before the model is used.
- Documentation sufficient for an independent party to understand and challenge the model.

These requirements apply to AI models without modification.

A traditional model encodes relationships specified in advance, so its assumptions are documented and can be inspected directly.

An AI model learns its relationships from data, so its assumptions are not stated but embedded across model parameters that cannot be read directly, and the same requirements become significantly harder to meet. Three properties drive this:

- **Opacity:** AI models' internal logic is not directly legible.
- **Data-dependent:** AI models' behaviour is determined by training data rather than by stated relationships.
- **Dynamic:** AI models' behaviour may change as input data shifts or as the model is retrained.

Development controls concentrate on the factors that determine the model's behaviour (i.e. training data, design choices, and observed performance under testing) rather than on a written specification.

The development record sets out the design rationale, the alternatives considered, and the testing evidence, to the standard that independent validation will require.

When developing AI models, specific consideration around Fairness, Explainability and, where the model is a frontier model accessed via a vendor, Frontier Model Configuration should be addressed.

AI- specific risks in development

 Data	 Model design	 Testing
AI models' behavior are determined by their training data, so weaknesses in the data become weaknesses in the model directly. The principal concerns are: • Representativeness - where data that does not represent the operating population produces poor or unfair performance on under-represented groups. • Data quality and lineage - including the integrity of labels produced by a prior process; and • Bias at source - where historical data reflects bias the model reproduces and may amplify. Response Evidence representativeness against the intended population, assess data quality and lineage, and address bias at source rather than deferring it to validation.	The model type should follow from the problem, the available data, and the interpretability the use case requires. The central judgement is the trade-off between performance and explainability: • More complex models can perform better but at the cost of transparency, justified only where the performance gain is material. The PRA has observed model-selection trade-offs made without adequate justification Response Benchmark the candidate model against a simpler, interpretable alternative, treat the added opacity as unwarranted where the improvement is marginal, and document the basis for the choice.	Testing assumptions that hold for traditional statistical models often do not carry across to AI models. These include: • Stationarity • Independent observations • Stable input distributions Response Extend developer testing beyond average performance to out-of-sample and out-of-time performance, stability in edge and corner cases, sensitivity to input changes, and resilience to adversarial manipulation and misuse; developers should know what must be reported to the model risk function.

Development

Fairness

Fairness should be addressed during development, not tested only at the end, and whether it is in scope is decided during risk classification:

- It is relevant where a model's decisions affect individuals or firms and there is an asymmetry of power, such as consumer lending and insurance access and pricing.
- Variables prohibited by law are excluded from the model's decisions, and the use of permitted but sensitive variables is a documented management judgement.
 - One of the most underestimated issues is the use of proxy variables: removing a protected variable does not remove the associated bias if it can be reconstructed from other inputs, so fairness by design means testing for proxy relationships.
 - Fairness also involves trade-offs that should be stated openly. Its definitions compete and removing a measured disparity costs accuracy and can itself cause harm, so any residual decision should be formally documented and approved by management.
 - In the EU, credit scoring and life and health insurance pricing are high-risk under the [EU AI Act \(Annex III\)](#).

Explainability

Explainability should be designed into the model rather than added afterwards; [SS1/23](#) names interpretability and transparency among the factors to be weighted for AI models.

- It trades against performance and depends on the use case and the audience, so the level is set at design time, driven by the model's risk tier and by who needs to understand its decisions.
- The framework should be candid about the limits of post-hoc techniques such as LIME and Shapley-value methods.

- Many assume features are independent and mislead when features are correlated, depend on parameter choices, or rely on surrogate models that do not reflect the complex model. The PRA made the same point (Observation 3, [PRA AI & ML Roundtable](#)), namely that explanations are a form of evidence and should be subject to challenge.
- Good practice is to use a simple interpretable model as a benchmark, to apply more than one technique whose weaknesses do not overlap, and to document the basis for the interpretability level chosen.

Development controls alone are insufficient for robust model governance but provide the foundation on which it can be built. Further to these controls is the requirement for independent model validation to provide essential challenge and assurance that the model is functioning within the scope assigned in the development stage.

Frontier model configuration

Where an AI model is a frontier model consumed through a vendor API rather than trained in-house, development risk does not disappear. It shifts to the decisions the firm still controls: which model and version is selected, how the system prompt and surrounding context are designed, and which sources feed retrieval. These are first-line development decisions and should be documented, version-controlled and tested to the same standard [SS1/23](#) expects of methodology choice, even though the underlying model itself was not built by the firm.

Model and version selection should be evidenced against a defined benchmark before deployment, and re-tested whenever the vendor changes the underlying model version, since a frontier model's behaviour can shift without a formal retraining event the firm initiates or controls. System prompts and retrieval sources should be treated as part of the model specification, not as informal configuration, because they materially determine what the model does.

Validation

The objective is to independently assess whether an AI model is fit for purpose, reliable, explainable, compliant, and appropriately controlled throughout its lifecycle.

Under [SS1/23 Principle 4](#), models are subject to independent validation that provides effective challenge; with the depth of validation proportionate to the model's materiality. These requirements apply equally to AI models without modification.

For AI models, the principle is unchanged, but the validation approach must change, and in one respect it changes fundamentally. An AI model's behaviour can change after deployment; validation cannot be a one-time exercise carried out at approval; for AI models it must be a continuous discipline with defined triggers for revalidation. This matters because validation is the principle most often under-prioritised in practice, and an AI model's behaviour is more likely than a traditional model to change in ways its initial validation did not anticipate.

Revalidation should not only be calendar-driven, but it should also be triggered by material changes in data, model performance, use case, customer population or control environment. In contrast, monitoring should be ongoing and continuous to reflect the nature of AI models and prevent model drift.

Independent validation of an AI model sets out the approach used, applies effective challenge across the relevant dimensions, and is supported by people with the competence to challenge the model effectively.

The validation approach determines both the dimensions covered and the skills required.

AI considerations in validation

 Conceptual soundness	 Data	 Accuracy and stability	 Explainability	 Bias
• Harder to assess, because the validator cannot read the model's logic from a specification. • Assessment shifts onto the design choices, the data, and observed behavior, and onto the prior question of whether the model, and this type of model, was the right tool for the problem.	• The data used to build the model is validated in its own right (in terms of representativeness, quality, leakage and label integrity) because, as in development, the data largely determines the model.	• Performance testing is central and accuracy is the first measure of performance against appropriate metrics, out-of-sample and out-of-time, and stability in edge and corner cases rather than only on average, since a model accurate on aggregate can be unreliable in the tails. • Benchmarking against a simpler model, and stress and scenario testing, complete the assessment.	• Where independent validation must be most sceptical: the validator produces and critically reviews explanations rather than accepting them at face value, uses more than one technique, and pays particular attention where techniques disagree or a simple and a complex model differ in the importance assigned to an input (Obs. 3).	• Applies effective challenge to the work done at development: the validator independently measures outcomes across the relevant groups and protected characteristics, repeats the proxy check, and tests whether the trade-offs made at development were reasonable and properly owned.

Validation

Model drift and revalidation

Under [SS1/23](#), validated models are subject to periodic revalidation along with ongoing monitoring to confirm that the model continues to perform as intended. Where AI models are subject to continuous improvement, there is scope for the model to “improve” beyond the ability to perform its original purpose. This is of key concern when considering the ability to perform full runtime reconstruction of an AI model, a key component in validating a traditional model.

An AI model’s behaviour can change as the data it receives shifts away from the data it was built on, or as the model is retrained, so its revalidation triggers must include drift and cumulative change and not only the calendar.

Traditional models are revalidated on a periodic, calendar-driven cycle, justified by the inability of the model to change without human intervention, which is not the case for an AI model, providing cause for additional monitoring. The PRA drew attention to cumulative change of this kind. The phenomena requiring monitoring include data drift, concept drift and model drift, and, for generative models, hallucinations, each assessed against the validated baseline.

Validation defines the triggers for revalidation and the standard against which a revalidated model is measured.

In short, monitoring detects when something has changed, and validation determines whether the changed model still meets the standard.

The operational mechanics (i.e. continuous, automated and event-driven monitoring of performance, data quality and drift, and the alert thresholds that prompt action) sit within ongoing monitoring and change management.

Agentic system validation

Model-level validation, however well executed, tests whether the frontier model performs as intended. It does not test whether the system built around that model behaves safely once it can plan, choose tools and act. Validating an agentic system is a distinct requirement, not an extension of model validation, and should cover:

- End-to-end scenario testing across realistic task sequences, not single-turn outputs alone.
- Adversarial testing, including attempts to manipulate the agent through its inputs, context or tools.
- Tool misuse testing, to establish what the agent does when a tool is unavailable, returns unexpected results, or is used outside its intended purpose.
- Testing for unexpected or unplanned actions the agent was not explicitly instructed to take.
- Continuous monitoring in production, not only point-in-time assessment at approval.
- Human override, to confirm escalation paths function as designed under load.
- Rollback and kill-switch testing, to confirm the system can be safely stopped and its actions reversed.

These tests sit alongside, not instead of, the model drift and revalidation triggers set out above. A firm that revalidates the frontier model on schedule but never tests whether the agent misuses a tool has addressed one layer of risk and left the rest unexamined.

While model validation provides assurance over the design, performance, and governance of internally developed AI models, organisations increasingly rely on externally sourced solutions to support business processes. As a result, it is equally important to establish appropriate oversight and risk management arrangements for third-party models, ensuring that risks are identified, assessed, and monitored throughout their lifecycle.

Validation

Third-party and vendor AI models

The use of third-party and vendor-provided AI models can enable the rapid adoption of advanced capabilities without the need for in-house development. However, these models may introduce additional risks relating to transparency, explainability, data handling, vendor dependency, and ongoing performance monitoring. As organisations often have limited visibility into the design and operation of externally developed models, appropriate due diligence, governance, and oversight must be applied throughout the model lifecycle to ensure that associated risks remain within the organisation's risk appetite and regulatory obligations.

At procurement

The established principle does not change:

- A third-party model must meet the same standard as one developed in-house, which requires the provider to justify the model's design, supply evidence of testing, and document it adequately.
- AI makes this harder to enforce, because vendors expose less of their models, bear less of the risk than the deploying firm, and often provide thinner documentation, and because the training data originates outside the firm, its representativeness and freedom from bias cannot be assumed.
- The firm should require testing results and documentation as a condition of use, run its own benchmarking and sensitivity analysis where these fall short, and evaluate the vendor's own data and fairness practices.

Under the [EU AI Act](#), a deployer that significantly modifies or integrates a general-purpose AI model can itself become a provider and inherit the associated obligations. Reliance on a small number of providers also raises concentration, controllability and contingency considerations.

The use of a third-party model does not transfer accountability. The firm may rely on vendor information, but it still needs to understand the use case, limitations, controls, monitoring arrangements, exit options and residual risk.

Frontier models introduce a sharper version of this dependency. A vendor can change a model's behaviour through a version update the firm did not request and does not control, with no formal retraining event to trigger revalidation on the firm's side. Firms should track model versions in use, monitor for vendor-announced changes, and re-benchmark performance on version change rather than assuming continuity. Where a small number of frontier model providers support many otherwise unrelated agentic use cases, the concentration risk already noted above compounds accordingly, and contingency arrangements should be assessed across the full set of dependent use cases, not case by case.

Validation

At validation

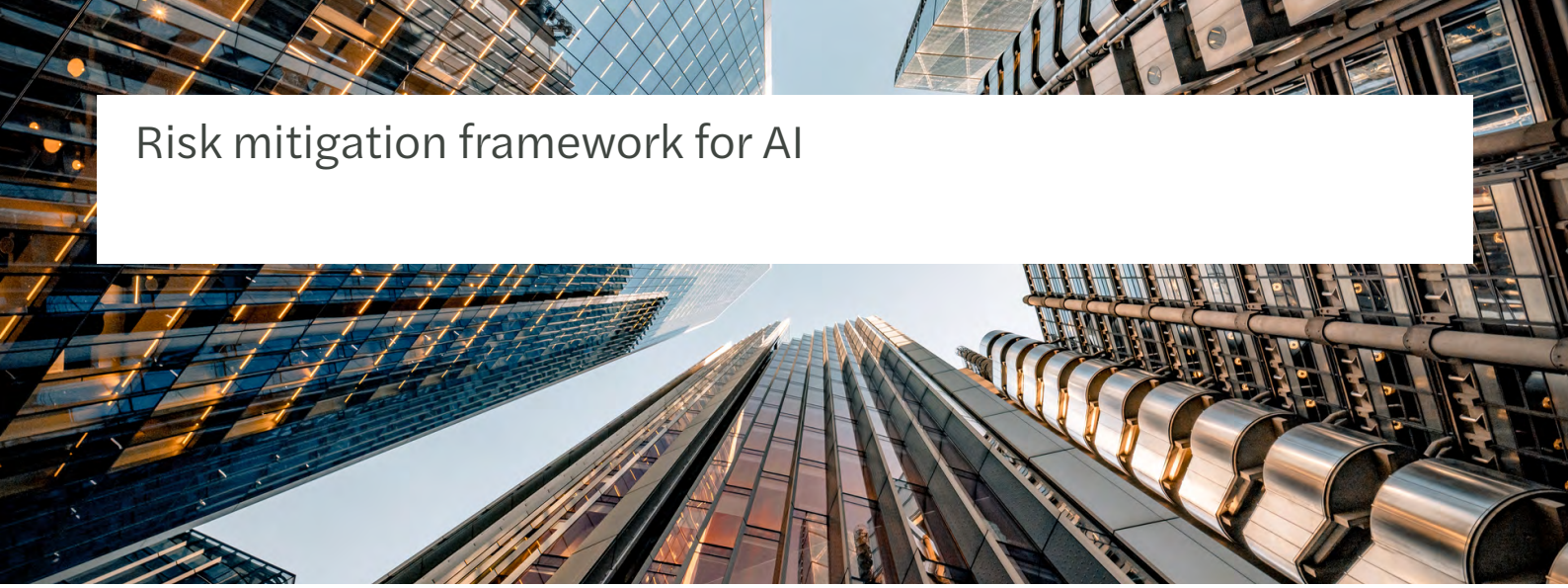
Validating a model the firm did not build raises a distinct challenge:

- The standard is the same, but the validator typically has less to work with.
- Vendor-supplied explanations carry an inherent conflict of interest, since the party providing them created the model and has an interest in its acceptance, and should only be relied upon where they are sufficiently comprehensive and independently credible.
- Where full access is not available, the validator compensates through independent benchmarking, sensitivity analysis and outcome testing, and documents the residual limitation rather than leaving reduced visibility unremarked.

While the challenges associated with third-party AI models are significant, they form only one part of a broader set of weaknesses observed across organisations adopting AI.

In practice, organisations should adapt existing model development standards rather than create entirely separate processes for AI models. The core principles of sound model development remain unchanged, but additional focus is required on training data quality, fairness, explainability, and reproducibility. Firms should ensure that key design decisions, model limitations and performance trade-offs are documented clearly enough to support effective independent challenge and regulatory scrutiny.

Forvis Mazars helps organisations strengthen AI model development frameworks by embedding AI-specific controls into existing MRM processes. We support clients in developing development standards, documenting model design decisions, assessing fairness and explainability considerations. This helps firms develop AI models that are both innovative and capable of withstanding independent challenge, validation and regulatory scrutiny.



Risk mitigation framework for AI

We need to build the guardrails to keep the AI models aligned with their original purpose. Effective AI risk management requires organisations to balance innovation with accountability. As companies adopt AI-driven solutions, they should identify, assess, and manage potential risks using an appropriate response strategy. Below are the four primary strategies for managing AI risk.

Mitigate

The company can introduce the safeguards to reduce the risk. Various tools to mitigate risks are discussed in the following section.

Transfer

Use of vendors or insurance may transfer elements of operational or financial exposure, but it does not transfer regulatory accountability or responsibility for customer outcomes.

Avoid

If the stakes are too high, the company should avoid or abandon such high-risk cases.

Accept

For low-stakes internal tasks, proceed where residual risk is acceptable if the consequences are minimal.

For firms seeking to manage AI risk, the task is often daunting. To support firms in translating risk management strategies into practice, a range of tools and techniques are available to aid organisations in identifying, assessing, monitoring and mitigating AI risk.

Risk mitigation framework for AI

Tools for risk mitigation

When mitigating key AI risks, firms must be aware of the various tools that are present in the market, before judiciously selecting the tools most effective for mitigating risk and validating their models. The graphic below details four effective tools firms may employ to mitigate risk:

The four tools illustrated below, human-in-the-loop, dynamic retraining, fallback mechanisms and prompt engineering, address risk at the level of the model's output. An agentic system introduces a fifth requirement: control over the system's actions. This includes action permissioning and approval gates for high-impact steps, sandboxed or scoped execution so an agent cannot reach beyond its intended tools, and tested kill-switch and rollback mechanisms so an action already taken can be halted and, where possible, reversed. Human-in-the-loop should itself be tiered by autonomy level, pre-action approval, post-hoc review, or continuous oversight, rather than treated as a single control applied uniformly.

Human-in-the-loop

For high-stakes decisions, ensure a human reviews or overrides AI suggestions. This can be automated (flag certain cases for review) or culture-driven (analysts double-check algorithms in real time).

Dynamic Retraining Pipelines

Trigger automated model retraining using fresh, validated data. Establish strict automated guardrails to ensure retraining pipelines do not introduce new biases or hallucinations.

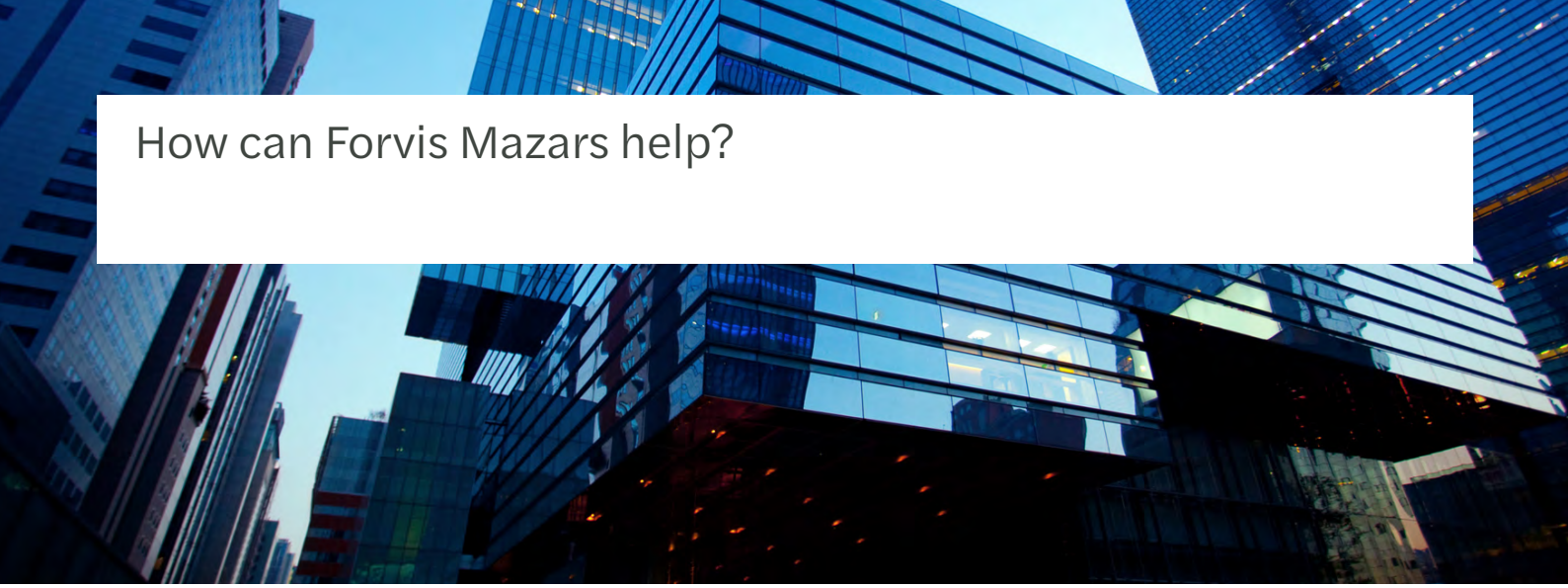


Fallback Mechanisms

Deploy alternative legacy models or business-logic "rules engines" to take over in the event of severe AI degradation or failure.

Prompt Engineering Adjustments

For Generative AI, apply automated prompt tuning or update guardrails and system instructions to address newly surfaced edge cases.



How can Forvis Mazars help?

Forvis Mazars helps organisations design and enhance Model Risk Management (MRM) frameworks that are proportionate, practical and aligned to regulatory expectations.

We support clients across the full model lifecycle, from model inventory development, risk tiering and governance design through to validation, monitoring and independent assurance. Drawing on experience of model governance, spreadsheet risk management and emerging AI controls, we help firms establish frameworks that enable innovation while maintaining appropriate oversight and accountability.

Our services include model inventory reviews, framework maturity assessments, policy and procedure development, model validation, internal audit support, and the implementation of controls to identify unregistered models and AI use cases. By combining risk, technology and regulatory expertise, we help organisations build sustainable MRM capabilities that enhance decision-making, strengthen compliance and reduce model-related risk.

This extends to agentic systems built on frontier models. We help firms apply an MRM classification assessment to agentic systems, decide whether they enter the model inventory or the AI system inventory, and design proportionate controls across the frontier model, context, agent harness, memory, tools and guardrails, not the model alone. Where firms are validating agentic systems for the first time, we support end-to-end scenario testing, adversarial and tool-misuse testing, and the design of override, rollback and kill-switch controls.

Contacts

Prasad Prabhakaran

FS Head of Applied AI

Prasad.Prabhakaran@mazars.co.uk

+44 (0) 748 399 1077

Becca Peet

Director, Insurance Risk Consulting

Becca.Peet@mazars.co.uk

+44 (0) 792 910 8177

Breda Griffin

Director, Banking Risk Consulting

Breda.Griffin@mazars.co.uk

+44 (0) 787 040 9159

Cameron Saycell

Manager, Insurance Risk Consulting

cameron.saycell@mazars.co.uk

+ 44 (0) 758 099 6846

Forvis Mazars is the brand name for the Forvis Mazars Global network (Forvis Mazars Global Limited) and its two independent members: Forvis Mazars, LLP in the United States and Forvis Mazars Group SC, an internationally integrated partnership operating in over 100 countries and territories. Forvis Mazars Global Limited is a UK private company limited by guarantee and does not provide any services to clients. Forvis Mazars LLP is the UK firm of Forvis Mazars Group.

© Forvis Mazars 2026. All rights reserved. 2609-10086