Financial Stability Institute | BIS

Occasional Paper
No 24

Managing explanations: how regulators can address AI explainability

by Fernando Pérez-Cruz, Jermy Prenio, Fernando Restoy, Jeffery Yong

September 2025

JEL classification: C60, G29, G38, O30

Keywords: artificial intelligence, machine learning, model risk management, risk governance

FSI Occasional Papers aim to contribute to international discussions on a wide range of topics of relevance to the financial industry and its regulation and supervision. The views expressed in them are solely those of the authors and do not necessarily reflect those of the BIS or the Basel-based standard-setting bodies.

This publication is available on the BIS website (www.bis.org). To contact the BIS Global Media and Public Relations team, please email media@bis.org. You can sign up for email alerts at www.bis.org/emailalerts.htm.

ISSN 1020-9999 (online)

Abstract

The increasing adoption of artificial intelligence (AI) by financial institutions is transforming their operations, risk management and customer interactions. Nevertheless, the limited explainability of complex AI models, particularly when used in critical business applications, poses significant challenges and issues for financial institutions and regulators. Explainability, or the extent to which a model's output can be explained to a human, is essential for transparency, accountability, regulatory compliance and consumer trust. Yet, complex AI models, such as deep learning and large language models (LLMs), are often difficult to explain. While there are existing explainability techniques that can help shed light on complex AI models' behaviour, these techniques have notable limitations, including inaccuracy, instability and susceptibility of misleading explanations.

Limited model explainability makes managing model risks challenging. Global standard-setting bodies have issued – mostly high-level – model risk management (MRM) requirements. However, only a few national financial authorities have issued specific guidance, and they tend to focus on models used for regulatory purposes. Many of these existing guidelines may not have been developed with advanced AI models in mind and do not explicitly mention the concept of model explainability. Rather, the concept is implicit in the provisions relating to governance, model development, documentation, validation, deployment, monitoring and independent review. It would be challenging for complex AI models to comply with these provisions. The use of third-party AI models would exacerbate these challenges.

As financial institutions expand their use of AI models to their critical business areas, it is imperative that financial authorities seek to foster sound MRM practices that are relevant in the context of AI. Ultimately, there may be a need to recognise trade-offs between explainability and model performance, so long as risks are properly assessed and effectively managed. Allowing the use of complex AI models with limited explainability but superior performance could enable financial institutions to better manage risks and enhance client experiences, provided adequate safeguards are introduced. For regulatory capital use cases, complex AI models may be restricted to certain risk categories and exposures or subject to output floors. Regulators must also invest in upskilling staff to evaluate AI models effectively, ensuring that financial institutions can harness AI's potential without compromising regulatory objectives.

Contents

Managing explanations: how regulators can address AI explainability issues

Section 1 – Introduction

Artificial Intelligence (AI) models are increasingly being used across all business activities of financial institutions, from internal operations to customer-facing applications. Crisanto et al (2024) and FSB (2024) highlight recent AI use cases in the financial sector, finding that most applications are for internal productivity-enhancing purposes. Financial institutions appear cautious regarding the use of AI for critical business applications, especially those involving customer interactions. Nevertheless, it is expected that the use of AI will become more prevalent, including in critical business areas, as firms seek to benefit from time and cost efficiencies, improved client services and enhanced regulatory compliance and risk management.

A key regulatory/supervisory concern is the explainability of AI models,[1] especially for critical business activities (eg customer-facing, core activities such as underwriting or determination of capital requirements).[2] While there is no commonly agreed definition of explainability, several organisations have defined this concept from their vantage points.[3] US regulatory agencies define explainability as "how an AI approach uses inputs to produce outputs".[4] This is the same definition referenced in FSB (2024). The OECD AI Principles, on the other hand, take a more customer-centric focus, defining explainability as enabling people affected by the outcome of an AI system to understand how it was arrived at by providing easy-to-understand information. IAIS (2025) defines meaningful explanations as providing understandable, transparent and relevant insights into how an AI system makes decisions or predictions. PRA (2023) describes explainability as the degree to which the workings of a model can be understood in non-technical terms.

The lack of explainability of the results of certain AI models can give rise to prudential concerns. FINMA (2024) points out that some AI model results cannot be understood, explained or reproduced and therefore cannot be critically assessed. The lack of explainability of certain AI models can also make it challenging for regulators to ascertain financial institutions' compliance with existing regulatory requirements surrounding the use of models, especially for critical business areas. IAIS (2025) highlights how model risk arising from the lack of explainability of complex AI models can result in unwarranted or unlawful trends (eg underpricing risks) going undetected, which can ultimately affect insurers' profitability and balance sheets.

Explainability is also important where AI models are used for calculating regulatory capital. When internal models were first allowed in the Basel Framework, the BCBS emphasised that "...a 'black box' that is not well understood by bank personnel does not provide confidence over the rating process or the

[1] See, for example, IIF-EY (2025), which shows that AI explainability is the top issue raised by financial institutions when engaging with regulators/supervisors on AI.

[2] For example, Intesa Sanpaolo has used machine learning to calculate regulatory capital for credit risk.

[3] A related – but often considered distinct – concept is interpretability. IBM considers interpretability to be focused on understanding the inner workings of an AI model, in contrast to explainability, which aims to provide reasons for a model's output.

[4] See OCC et al (2021). OCC (2021) defines AI explainability as the extent to which AI decisioning processes and outcomes are reasonably understood by bank personnel.

estimation of PDs".[5] Basel III tempered the use of such models by removing the use of the advanced options for certain risk categories and introducing input and output floors.[6]

The lack of AI model explainability could potentially contribute to heightened systemic risk.[7] FSB (2024) identifies model risk, data quality and governance as sources of AI-related vulnerabilities that can pose risks to financial stability. The report points out that the complexity and limited explainability of certain AI models and the difficulty in assessing data quality could increase model risks for firms that lack robust AI governance.

Explainable AI model outputs are also important from a consumer protection perspective, to avoid discriminatory decisions. IAIS (2025) points out that explainability and transparency are key to building trust and holding firms accountable for risks posed to consumers, such as unlawful discrimination. The complexity of certain AI models can make it challenging to explain model outcomes to consumers and can lead to increased risk that biases go undetected.

From the perspective of financial institutions, lack of explainability constitutes an obstacle for the adoption and deployment of AI models. Financial institutions would be wary of using AI models if they do not know how these models work. The use of AI models that lack explainability amplifies financial institutions' model risk. As such, overcoming the explainability challenge is important to avoid missed opportunities in terms of AI use cases that can enhance customer experience, regulatory compliance, risk management and operational efficiency.

Therefore, supervisors generally expect firms to be able to explain AI models that are used for critical activities or to inform decision-making in order to ascertain whether the model outputs are appropriate and to hold firms accountable. Overarchingly, a supervisor is unlikely to trust the results of an AI model if its results cannot be understood.

As mentioned, AI methodologies may help financial institutions to increase the efficiency of their operations, improve risk management and provide clients with more and better services. Thus, some supervisors have started recognising that overly stringent explainability requirements could impede socially desirable innovation. For example, some authorities now implicitly allow or explicitly recognise that some financial institutions may be using AI models for regulatory capital purposes.[8] This may be meant to encourage innovation and the development of AI use more generally. At the very least, it indicates the need for financial authorities to critically review their current guidance for the use of models by regulated institutions, taking into account the explainability challenges that AI models can entail.

There are existing international standards and jurisdictional regulatory requirements on model risk management (MRM), some of which already explicitly cover or implicitly allude to explainability issues.[9] Nevertheless, these requirements are typically high-level and may not capture the specificities of AI models. As such, a clearer articulation of the concept of explainability with regards to AI models would be helpful to supplement existing MRM requirements.

The aim of this paper is to describe current regulatory guidance on MRM; discuss existing challenges in applying it to AI models, especially in the context of limited explainability; and put forward considerations for addressing some of these challenges. For that purpose, the paper reviews existing

[5] BCBS (2001).

[6] See BCBS (2017).

[7] See, for example, Daníelsson et al (2022).

[8] See, for example, EBA (2023) and Bank of England and Financial Conduct Authority (2024).

[9] IAIS (2025) stands out as an example of international guidance to insurance supervisors on how to supervise the use of AI by insurers, also covering explainability expectations.

guidance on MRM, identifies elements related to model explainability, analyses the extent to which that guidance could be met by AI models and discusses possible elements of improvement in the current policy setup.[10]

The structure of the paper is as follows: Section 2 discusses MRM requirements that may be affected by lack of AI explainability. Section 3 outlines the different AI explainability methodologies and the challenges of implementing them in the context of more complex AI models. Section 4 describes potential adjustments to existing MRM requirements to address these challenges. Section 5 concludes.

Section 2 – MRM and explainability

Global standard-setting bodies (SSBs) already provide some high-level requirements on the use of models by financial institutions. The Basel Core Principles (BCPs), specifically BCP 15 (Risk management process) essential criterion 6, state that banks using risk models must comply with supervisory standards on model use, including the conduct of independent validation and testing of models.[11] The Insurance Core Principles (ICPs), specifically ICP 16 (Enterprise risk management for solvency purposes), address the use of models for risk measurement, including for measuring technical provisions.[12] There are also other pertinent issuances by the BCBS related to model use, such as the principles for effective risk data aggregation and risk reporting, which call for accurate and reliable generation of risk data by banks, and the principles on stress testing, which provide that models should be fit for purpose and subject to challenge and regular review.[13]

SSBs have also issued more detailed requirements on the use of models for regulatory capital purposes. The IAIS provides detailed requirements through its ICP 17 (Capital adequacy), while the BCBS has risk-specific guidance, particularly the internal ratings-based (IRB) approach for credit risk and the internal models approach (IMA) for market risk.[14] All require effective governance and controls for the use of internal models and assign ultimate responsibility to the board and senior management for understanding the consequences and limitations of model outputs. They also require financial institutions to ascertain that model methodology and assumptions are conceptually sound, are fit for the intended purposes and have good predictive power. Another common requirement is documentation of model design and assumptions.

More recently, IAIS (2025) elaborates on how existing ICPs apply in the context of insurers' use of AI. It covers areas of governance and risk management that have been identified as requiring particular attention when deploying AI. It also covers both technical aspects, such as data governance and model validation, and other activities that aim to support supervisors and insurers in managing risks that are introduced or enhanced by AI. Box 1 outlines how the ICPs can be applied to address the explainability of AI models.

[10] IIF-EY (2025) indicates that the top issue discussed between regulators and firms in relation to AI is explainability/the black box nature of certain AI algorithms.

[11] See BCBS (2024a).

[12] See IAIS (2024).

[13] See BCBS (2013) and BCBS (2018).

[14] See BCBS (2022) and BCBS (2024b).

Box 1

IAIS Application Paper on the supervision of artificial intelligence – explainability-related guidance

- ICP 7 (Corporate governance) – There should be clear accountability within an insurer for setting expectations on AI systems so that the output generated is explainable, fair and unbiased.
- ICP 8 (Risk management and internal controls) – Insurers should have effective risk management and internal controls to minimise the risk that AI systems have an adverse impact on their financial soundness.
- ICP 19 (Conduct of business) – The transparency and explainability of claims decisions and claims dispute resolution influenced by AI systems are especially important.
- ICP 8 (Risk management and internal controls) and ICP 19 (Conduct of business) –
    - Supervisors should require insurers to meaningfully explain the outcomes of the AI systems that they use, especially in use cases that have a material impact on consumers or solvency or those used to satisfy legal requirements.
    - Insurers could restrict deployment of AI systems to those that are simple and explainable or restrict the use of complex AI systems to challenging and fine-tuning more traditional mathematical models.
    - The deployment of complex AI systems could be conditional on the accompanying deployment of explainability tools such as Shapley values or LIME, which can be employed to illustrate the influence of different variables on AI outcomes, enhancing transparency and trust, with recognition of the limitations of such techniques.
    - Where the risks from the AI system are high and/or the tools used to explain the model themselves have limitations, insurers could instead consider alternative simpler models.
    - Where an AI system cannot provide sufficient confidence under new or unexpected conditions, insurers should ensure it can fail safely or escalate to human intervention.
    - For highly complex AI systems with which it is not possible to achieve the desired level of explainability, insurers should consider adopting and documenting complementary governance measures such as the use of guardrails (eg enhanced data management) and human oversight.
    - The level and depth of explanation of an AI model should be tailored to the different stakeholders (eg auditors and supervisors will require more comprehensive technical information compared with policyholders).

Source: IAIS (2025).

At the national level, there are only a few financial authorities with MRM guidelines in place. Building on the work by SSBs, those authorities – with either specific MRM guidelines or general risk management guidelines that touch on model use – usually focus on models that are used for calculating regulatory capital. Table 1 lists the MRM guidelines reviewed for this paper.[15]

[15] Guidelines issued by authorities that simply transpose SSBs' standards on the use of models for regulatory capital calculation (eg *ECB guide to internal models*) are not included here.

MRM guidelines

Table 1

	Issuing authority	Title of document	Month/year issued
Canada	Office of the Superintendent of Financial Institutions (OSFI)	Draft guideline E-23 – Model risk management*	November 2023
Japan	Financial Services Agency of Japan (FSA)	Principles for model risk management	November 2021
United Arab Emirates	Central Bank of the United Arab Emirates (CBUAE)	Model management standards	November 2022
United Kingdom	Prudential Regulation Authority (PRA)	Model risk management principles for banks	May 2023
United States	Federal Reserve Board/Office of the Comptroller of the Currency (FRB/OCC)	Supervisory guidance on model risk management	April 2011
	OCC	Model risk management (Comptroller's Handbook)**	August 2021

Links to the documents are provided in the References section.

*Original guideline was issued in 2017. Consultation on the draft guideline closed in March 2024 but no final version has been issued yet.

**The handbook aligns with the FRB/OCC 2011 guidance but updates it to cover AI-related issues.

MRM guidelines have common elements. All guidelines cover governance and oversight, model development and documentation, model validation and implementation, monitoring and maintenance. These guidelines also require the assessment of model risk in order to allow for a risk-based approach to the application of the MRM requirements. Moreover, all MRM guidelines cover the management of risks when using third-party models. A couple of the guidelines explicitly mention the issue of model explainability.

While the concept of model explainability is not explicitly mentioned in many of the existing MRM guidelines, it is implicit in many of the provisions contained in those guidelines.

- Governance: MRM guidelines emphasise the role of the board and senior management in ensuring that their firm's MRM framework is effectively implemented. In addition, many guidelines also expect them to have sufficient understanding of the models and provide effective challenge.
- Model development and documentation: MRM guidelines require that model documentation be transparent[16] and include, among others, the theory, assumptions, logic, specifications and limitations of the model. Some guidelines also require that the methodology be clear and well articulated to all stakeholders (CBUAE) and that the merits and limitations of the model be understood and communicated to model users and other stakeholders (PRA).
- Model validation: MRM guidelines require model validation to assess the suitability and conceptual soundness of the model. In addition, this assessment should be done by model validators that are independent from the model development process. One guideline (OSFI)

[16] Prenio and Yong (2021) explains how transparency is a precondition for enabling supervisory assessment of AI models. Transparency of AI models depends, among other factors, on whether they are built in-house, open or closed source, or proprietary applications developed by third-party service providers, including big techs. The US Federal Office for Information Security (2024) emphasises that transparency of AI systems should cover the entire life cycle of an AI system and its ecosystem. Regulators without a full view of the AI ecosystem may not be in a position to properly assess the explainability of an AI model.

explicitly expects model validation to ensure that the model is understandable to relevant stakeholders.

- Deployment and ongoing monitoring: One guideline (FRB/OCC) provides that business areas using the model should be able to question the model's assumptions if model outputs do not seem to make sense. Another guideline (CBUAE) mentions that the objective of the monitoring process is to assess whether changes in the operating environment have had an impact on the performance, stability, key assumptions and/or reliability of the model.
- Independent review or internal audit: In general, MRM guidelines only expect internal audit to assess the effectiveness of the implementation of the MRM framework. However, some guidelines (BCBS (2022, 2024b), PRA (2023)) also specify the need for independent review in order to assess the appropriateness and soundness of models.

The requirement to assess the riskiness of a model in order to allow for a risk-based application of MRM requirements exacerbates the implementation challenges. The factors that firms need to consider in assessing the riskiness of a model include materiality of use, complexity of the model,[17] uncertainty about inputs, approaches and assumptions, and potential customer impact. The riskier the model, the more frequent and intensive the application of the MRM requirements. As such, based on these factors, AI models that lack explainability and which are used in high-risk areas will likely be given a high-risk rating. Hence, somewhat ironically, the MRM guidelines would make explainability more relevant for models that lack it.

The use of third-party models also heightens the challenges arising from lack of explainability. Bank of England and Financial Conduct Authority (2024) finds that half of survey respondents reported having only a partial understanding of the AI technologies they use due to the use of third-party models. In general, firms are expected to comply with MRM requirements even for models provided by third parties. Some guidelines recognise the challenge facing firms that typically do not have full information about third-party models. The FRB/OCC and FSA guidelines allow validation work to be adjusted as a result. The former provides that banks may have to rely more on sensitivity analysis and benchmarking. The latter concedes that validation may need to be based on the best available information. The CBUAE guidelines, on the other hand, state that if a third-party model is not fully understood by the firm, then it must not be considered fit for purpose.

One aspect of MRM that is not explicitly covered in existing guidelines relates to firms' responsibility towards customers that are affected by model outcomes.[18] The discussion so far focuses on how the issue of explainability might affect the implementation of mostly "internal-facing" MRM requirements. However, recent AI policy issuances highlight the concepts of external transparency, external accountability and procedural fairness.[19] These concepts involve requirements for firms to inform customers when they are interacting with AI and about the use and consequences of AI-driven decisions that affect them, as well as to provide simple explanations of how the AI model works and contributes to the decision and channels for complaints and redress. These are important elements that should be covered by MRM guideless since they could result in significant reputational risk. At the same time, like

[17] In the PRA MRM guidelines, lack of explainability is explicitly mentioned as one of the factors that drives model complexity. OCC (2021) also states that examiners should assess whether model ratings take explainability into account.

[18] The MAS FEAT Principles, for example, state that financial institutions that use AI should provide data subjects (eg prospective financial customers) with channels for inquiring about, submitting appeals for and requesting reviews of AI-driven decisions that affect them. The EU AI Act has similar provisions, while Article 22 of the EU GDPR states that "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her".

[19] See, for example, Prenio and Yong (2021) and Crisanto et al (2024).

the internal-facing requirements mentioned above, implementation of these external-facing requirements would also be constrained by lack of explainability.

Section 3 – Challenges in enforcing explainability requirements in the context of AI

Firms may find it challenging to meet existing regulatory requirements on explainability of AI models.[20] Advanced AI models such as deep neural networks are difficult to explain due to their many parameters and over-parametrisation (ie more parameters than data points). Unlike simpler models such as linear regression, which have explainable mathematical relationships that can show how the data inputs result in the model outputs, advanced AI models are complex because they involve large volumes of non-linear reasoning. This black box nature of advanced AI models can obscure the computation behind model outputs. While advanced AI models can offer superior predictive performance, they often come at the expense of explainability, leading to concerns around accountability, fairness and ethics.[21]

The construction of large language models (LLMs) renders their functionality more complex than that of other AI models. LLMs are trained on vast data sets of considerable scale (comparable with the size of the internet), and their outputs are derived from intricate interactions involving billions of parameters. These models generate predictions for the next token based on probabilities, meaning that a random number draw determines the subsequent word among the most likely candidates. Consequently, the model's output may vary even when the input remains unchanged. In certain extreme scenarios, an unlikely draw can result in the model producing outputs referred to as "hallucinations". This probabilistic nature underpins the model's ability to address complex queries effectively, but it also introduces potential adverse effects. Furthermore, the training data and processes employed in developing LLMs are often opaque. Even in the case of models with open-access (ie publicly available) weights, it remains challenging to assess their future behaviour and delineate their capabilities and limitations.

3.1 Existing methods for explaining AI models and potential limitations

In most policy discussions, the term "explainability" is used, whereas in most academic literature, the term "interpretability" is used. While on the surface these terms appear to be interchangeable, they can convey different meanings. Clarifying the definitions is important so that the policy discussion is clear on the intended policy/regulatory objectives. This paper adopts the following understanding, based on a review of a selection of academic papers,[22] in order to properly frame the subsequent paragraphs:

[20] De Nederlandsche Bank and Dutch Authority for the Financial Markets (2024) highlights how AI models might not meet certain IRB requirements, for example the requirement for projected results to be "plausible and intuitive" – AI can sometimes produce results that are not intuitive but nevertheless can turn out to provide a better estimate of credit exposure than traditional models.

[21] EBA (2023) highlights the trade-off between explainability and model performance. While more complex models may yield better performance, they are more difficult to explain or comprehend.

[22] See Retzlaff et al (2024), Gilpin et al (2019) and Doshi-Velez and Kim (2017).

- Explainability refers to the extent to which a model's output can be explained to a human (it answers the questions "why did the model produce this output?" or "why did the model recommend accepting this credit application?").[23]
- Interpretability refers to the extent to which the inner workings of an AI model can be understood by a human (it answers the questions "how did the AI model arrive at this output?" or "how did this AI model determine that this property should not be insured?").[24]

These concepts are interlinked.[25] Inherently interpretable models are explainable, but the reverse is not true. One can "interpret" a model by describing what it will predict given an input. In practice, this can be achieved by assessing how the model is trained using training and testing data sets and tracing how the underlying algorithm processes new data input into the model output. For some complex models this may not be possible, but one can "explain" the model by describing the model's behaviour. This requires reasoning and justification for why the model made certain decisions.[26]

It is important to recognise the non-binary nature of these concepts, ie it is difficult to say conclusively or categorically that an AI model is or is not explainable. When a supervisor assesses whether an AI model is explainable, the conclusion may not always be "yes" or "no", but rather the conclusion could be "yes, to a certain extent". The threshold of acceptance is dependent on the context and the requirements of the assessor.

Certain AI models are inherently interpretable (also called a priori models), for example:

- Decision trees:[27] A model is explained by following the tree structure that follows an if-then-else-rules approach.
- Generalised additive model:[28] A model that explains the relationships between input variables and predicted output.

Nevertheless, there exist black box models that are inherently not interpretable because of their complexity, non-linearities and the use of a large number of parameters, making it impossible to connect the model output back to understandable rules, patterns or inputs.[29] To contribute to the explainability of these models, post hoc techniques can be used to analyse a black box model after it has made

[23] See Section 1 for a collection of definitions of explainability in the financial sector. Other useful references include European Data Protection Supervisor (2023).

[24] In some literature, explainability and interpretability are used interchangeably, but this paper finds it useful to distinguish between these concepts for regulatory purposes, as they may have different implications for financial institutions. Examples of academic definitions of interpretability are Gilpin et al (2019) and Lipton (2017). Lipton (2017) illustrates how the concept of interpretability is not straightforward and identifies features of an AI model that confer interpretability, such as transparency. It is acknowledged that there are other interpretations of explainability and interpretability, for example NIST (2023) takes an opposite view, stating that explainability answers the question of how a decision was made in the system, while interpretability answers the question of why a decision was made by the system and its meaning or context to the user.

[25] See Gilpin et al (2019). Thampi (2022) states that explainability requires interpretability as a building block.

[26] This paper focuses on the explainability of AI models that are not interpretable, because such models pose regulatory challenges. Per the definitions above, models that are interpretable should also be explainable and, thus, should pose fewer issues to regulators.

[27] See Molnar (2020).

[28] See Retzlaff et al (2024) and Saleem et al (2022).

[29] See Retzlaff et al (2024). European Data Protection Supervisor (2023) offers a similar classification but uses a different terminology: "white box" (self-interpretable models) versus "black box" (post hoc explanations). There are other categorisations of explainability techniques, for example Thomas (2024) summarises four main categories of general AI explainability techniques for GenAI models: feature-based, sample-based, mechanistic and probing-based.

predictions/delivered an output.[30] Post hoc techniques can be further broken down in terms of global and local explainability. Global explainability refers to the ability to explain the overall functioning of a model by capturing patterns, general trends and insights that apply broadly to its behaviour, while local explainability identifies specific inputs that drive a specific output.[31] The former is more of a concern for model developers, validators and regulators, while the latter is more of a concern for individual customers seeking explanations in use cases like credit decisions.

Post hoc techniques include:

- SHapley Additive exPlanations (SHAP) method,[32] which attributes a model's prediction to individual input features/factors;[33]
- Local Interpretable Model-agnostic Explanations (LIME) method,[34] which explains the most significant features that influence a prediction by fitting a simpler model using data that are slightly altered from the original data point; and
- Counterfactual explanation,[35] which explains a model output by identifying the smallest change to the features that would change the prediction/output.

Post hoc techniques are useful for overcoming the cognitive limits of humans, which may restrict the extent to which an AI model can be understood. Candelon et al (2023) refers to the cognitive load theory which suggests that humans can only understand up to about seven rules or nodes, making it virtually impossible for humans to fully understand decisions made by sophisticated AI systems.[36] In other words, while a firm may document how an AI model works, human supervisors may not completely understand. IMDA and PDPC (2020) points out that technical explainability may not always be sufficient, especially if the target audience is the average consumer. Providing laypersons with counterfactuals (eg "your mortgage would have been approved if your average debt was 15% lower") could be more appropriate in such cases.[37]

[30] Open source explainability algorithms are available to generate post hoc explanations; see Github.

[31] See, for example, MAS (2024).

[32] See Lundberg and Lee (2017). For example, for an AI model that is used for credit underwriting, SHAP can explain the extent to which each "feature" or risk factor (income, credit history etc) contributes to the model output (underwriting decision). Davis et al (2022) provides an example of how SHAP and LIME explained the credit assessment output of selected machine learning models. Bank of England and Financial Conduct Authority (2024) finds that feature importance and SHAP are the most widely used by surveyed financial institutions in the United Kingdom. Visualisation techniques might also help to facilitate explainability. Goldstein et al (2015) proposes individual conditional expectation (ICE) plots, which display per-instance prediction-feature curves to reveal heterogeneity and interactions that average partial dependence can obscure, along with a simple visual test for additivity. Apley and Zhu (2020) introduces accumulated local effects (ALE) plots, which are an efficient way to visualise feature effects by integrating local changes in predictions.

[33] IAIS (2025) provides examples of explainability techniques for selected AI use cases by insurers. For example, SHAP can be used to explain to consumers: "Your premium was influenced primarily by your driving history (40% impact), vehicle type (30% impact) and location (20% impact)."

[34] See Ribeiro et al (2016). For example, to explain an AI model's credit underwriting recommendation, the value of one or several risk factors is changed slightly and the model is re-run with these adjusted input values. Higher weights are assigned to risk factors that produce outputs that are more similar to the original data input, thus illustrating how the model arrives at the recommendation. Ribeiro et al (2018) provides an improved version of LIME called anchors.

[35] See Dandl et al (2020), which provides an example of a counterfactual statement: "You were denied a loan because your annual income was GBP 30,000. If your income had been GBP 45,000, you would have been offered a loan." EBA (2023) lists several measures used by surveyed banks to explain machine learning techniques; such measures include counterfactual explanations.

[36] The way in which explanations are delivered may influence the comprehension of the intended audience. Pearl and Mackenzie (2018) opines that humans are more attuned to causal narratives than abstract statistical concepts.

[37] See Russell et al (2018).

It is worth noting that these explainability techniques are not mutually exclusive and there are pros and cons for each method.[38] From a supervisory perspective, understanding the limitations of explainability techniques is important (Table 2 provides a non-exhaustive list of the limitations of explainability techniques).

Limitations of explainability techniques

Table 2

Limitation	Description
Inaccuracy	Explanations may not faithfully represent an AI model's actual decision.[1]
Instability and sensitivity	Small changes to data input can lead to drastically different explanations.[2]
Inability to generalise	Explanations may not hold true when generalised to a broader population/data set.[3]
Non-existence of ground truth	There are no universally accepted metrics to assess the correctness or completeness of explanations.[4]
Misleading interpretations	Misleading explanations can appear plausible.[5]

[1] Rudin (2019) explains how post hoc techniques like LIME or SHAP can be unfaithful to the underlying AI models. [2] Alvarez-Melis and Jaakkola (2018) explains how model-agnostic perturbation-based explainability techniques are more prone to instability than are gradient-based techniques. [3] Molnar et al (2020) explains that, even for global explainability techniques, under- or overfitting can result in poor model generalisation. [4] Bordt et al (2022) explains how post hoc techniques are not effective in adversarial contexts due to their inherent ambiguity; susceptibility to manipulation; and inability to provide unique, truthful reasons for algorithmic decisions. [5] Lakkaraju and Bastani (2020) shows how black box models can be manipulated to provide misleading explanations.

New explainability techniques[39] are being developed and existing methods improved,[40] all of which may prove useful in the future to assist financial sector regulators in better understanding AI models used by financial institutions.

3.2 Challenges in applying existing requirements

An overarching MRM requirement is for an AI model to be explainable with regard to how it arrived at an outcome.[41] If interpreted strictly, this requirement may be challenging for certain AI models such as deep learning methods. For example, deep neural networks consist of multiple hidden layers with thousands of parameters that interact in a non-linear way. It is not possible to conclusively attribute the model output to combinations of the data input. Complex models such as GPTs that use billions of parameters[42] are another example of how existing explainability requirements may not be fit for purpose.

[38] Retzlaff et al (2024) outlines criteria that can be used to compare explainability methods. Alonso-Robisco and Carbó (2025) provides an example of how to assess the relative reliability of SHAP and Permutation Feature Importance in explaining the credit decision predictions of selected AI models. In the context of banking supervision, SHAP can provide an answer to the question "which factors contributed most to the credit score of applicants in general?". Buckmann and Joseph (2023) introduces a three-step workflow to make machine learning forecasts interpretable and reveal economically meaningful non-linearities.

[39] See, for example, Wu et al (2024).

[40] See Dhurandhar et al (2023). Alonso-Robisco et al (2025) proposes constraining a model's internal logic during training in order to improve both the credit prediction performance of the model and its interpretability.

[41] See OSFI (2023).

[42] Heaven (2023) reports that OpenAI's GPT-3 has 175 billion parameters, which are values in a neural network that are adjusted during the training process.

In addition, explainability requirements may need to be tailored to the target audience, for example senior management, consumers or regulators.[43] Yet, most existing requirements do not make such differentiation. FINMA (2024) mentions that, where decisions had to be justified to investors, clients, employees, the supervisory authority or the audit firm, FINMA assessed the explainability of the applications in greater depth. The assessment includes understanding the drivers of the applications or the behaviour under different conditions in order to be able to assess the plausibility and robustness of the results.

Some MRM requirements specify model change processes that firms need to follow; however, it is unclear what constitutes a change when it comes to AI models. It would be necessary to specify what constitutes a model change for AI models. Bank of England (2025) notes the potential for dynamism in complex AI models that get automatically updated as new data are available.

The use of AI models provided by third-party providers poses multiple challenges in observing MRM requirements.[44] In certain jurisdictions, firms are required to engage independent third parties to undertake model validation. For proprietary models licensed from third-party providers, such independent model validation may be challenging as providers may be reluctant to explain how their model works or is trained for proprietary reasons. This situation is observed in the area of LLMs; firms generally have neither the financial nor the computational resources to develop their own models.[45]

Different types of AI model may present different levels of challenges in observing MRM requirements. MAS (2024) highlights that, compared with conventional AI, which is typically used by banks for specific use cases for which the AI models have been trained, generative AI (GenAI) is more general-purpose in nature and can be used in a wider range of use cases in the bank. In some of these newer use cases, it might be difficult to find readily available ground truths for evaluating and testing GenAI models.

The lack of established or globally accepted explainability methods, especially for newer forms of AI models, is a barrier to meeting MRM guidelines. MAS (2024) observes that there is a general lack of established methods for explaining GenAI outputs and assessing their fairness.

[43] OSFI and Global Risk Institute (2023) outlines four factors to consider in determining the appropriate level of explainability: what needs to be explained, target audience, materiality of the use case and complexity of the model. IAIS (2025) discusses tailoring explanations of AI outputs to different stakeholders. Davis et al (2022) analyses the explainability of machine learning models of consumer credit risk for various stakeholders: loan companies, regulators, loan applicants and data scientists.

[44] MAS (2024) states that the lack of transparency of external model providers may contribute to challenges in understanding and explaining the outputs and behaviour of GenAI.

[45] In the EU, this is somehow mitigated by the Code of Practice for model developers/providers. The first chapter of the code relates to transparency and includes a model documentation form that outlines the information that model developers/providers need to share in order to ensure sufficient transparency.

Box 2

Explainability of large language models

Large language models (LLMs) are increasingly being used by financial institutions across many activities, from information search and summarisation to generation of client or management reports. LLMs are very powerful because their underlying foundation models are trained on massive data sets, so much so that people often think only about LLMs when talking about AI.

LLMs' impressive ability to respond to any question has attracted admiration and confidence from a growing number of users, despite the fact that they do not know exactly how the models work.

Yet, explaining and understanding LLMs is an extremely complex task due to the many parameters and non-linear relationships involved. Ameisen et al (2025) offers a glimpse into Anthropic's LLM, Claude 3.5 Haiku, using attribution graphs, which aim to partially trace the chain of intermediate steps that the model uses to transform a specific input prompt into an output response.

Another approach to understanding what is going on in an LLM is through chain-of-thought (CoT) prompting, first coined by Wei et al (2023). Originally, CoT referred to humans detailing their reasoning process in the prompt to help earlier-generation models arrive at better answers. Humans would provide the model with the reasoning steps, which would then be used by the model as an example of how to produce the reasoning for other tasks. These methods served as a guide to improve the model's performance by mimicking structured human reasoning. These prompts are interpretable; however, how LLMs use them is not.

With the advent of newer-generation reasoning models, the focus has shifted. These newer models aim to reproduce human-like reasoning patterns more directly.① Importantly, while the explanations generated by these models may appear to mimic how humans reason, they do not necessarily reflect how the models actually arrive at their conclusions. In other words, producing an appealing explanation does not guarantee that it represents the model's internal reasoning process.

As more firms develop AI applications based on LLMs, depending on the use case, it may become more of a regulatory concern if they are unable to fully explain how the applications work. This problem will not go away any time soon given that firms may never know how the foundation models were trained, including the data used.

① OpenAI states that its reasoning models are LLMs trained with reinforcement learning to perform reasoning, ie they produce an internal chain of thought before responding to a user. The models break down a prompt and consider multiple approaches to generating a response.

Section 4 – Potential adjustments to MRM guidelines

Given the discussions above, authorities may need to review existing MRM guidelines and determine whether new guidelines or adjustments to existing ones are required to preserve their relevance in the age of AI. As financial institutions expand their use of AI models for different functions and business areas, financial authorities may need to provide guidance on the use of models beyond regulatory capital purposes. Clear AI guidance provides regulatory certainty and can help support the use of models in a safe and sound manner, while safeguarding consumer interests. MRM guidelines do not need to be technology-specific, but they do need to be regularly updated to reflect issues associated with the use of newer technologies. This is especially important as financial institutions' use of models utilising advanced technologies becomes more prevalent.

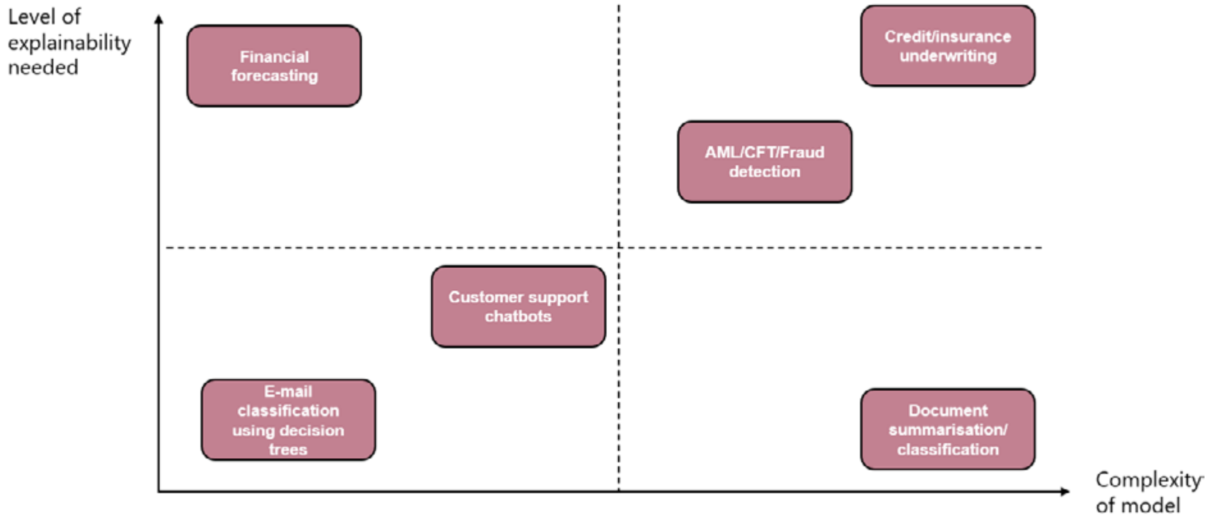
In principle, where AI models are used for making decisions in critical activities, MRM guidelines could require financial institutions to use inherently explainable AI models[46] or at least adopt sufficiently

[46] See Alonso-Robisco and Carbó (2025).

informative explainability techniques for black box models. Yet, for complex models, using just one existing explainability method may not be fully informative. As such, it may be necessary to require a suite of methods to address the differing needs of stakeholders. Various stocktakes, surveys or thematic reviews conducted by both the private and public sectors show that financial institutions are using several explainability techniques to better understand complex AI models.[47]

Stylised illustration of relative complexity of AI models for selected use cases and the corresponding relative level of explainability required

Graph 1


Level of explainability needed
Financial forecasting
Credit/insurance underwriting
AML/CFT/Fraud detection
Customer support chatbots
E-mail classification using decision trees
Document summarisation/ classification
Complexity of model

MRM guidelines may need to require financial institutions to establish acceptable explainability standards for relevant use cases. For example, OCC (2021) requires that, for models that may be difficult to evaluate for conceptual soundness, a determination should be made as to whether the level of explainability is appropriate to the specific model's use. Firms' model documentation, including validation reports, should include information on how the explainability of AI models used in high-impact cases has been addressed. The information provided should enable supervisors to assess why a model has made specific predictions. Graph 1 illustrates the different levels of explainability that a supervisor might require, depending on the criticality of a specific use case and the level of complexity of the model.

More generally, tailoring regulatory explainability requirements to the different levels of riskiness of AI use cases could be considered. Tiering or rating of models based on their riskiness, which is already required in existing MRM guidelines, can be used to identify use cases for which explainability is more relevant. Banks and other financial institutions and authorities are already starting to do this.[48] Typically,

[47] See, for example, EBA (2023), Bank of England and Financial Conduct Authority (2024), MAS (2024) and IIF-EY (2025).

[48] See, for example, OCC (2021), MAS (2024), FINMA (2024) and EIOPA (2025).

AI use cases that are subject to higher standards of explainability are those that result in decisions that likely require explanations to external stakeholders, such as regulators and customers, or those for which inaccurate model outputs can have significant (financial) implications.

There should be explicit recognition of possible trade-offs between explainability and model performance. Complex models may bring improved performance, which should be weighed against the consequences of insufficient explainability.[49] A possible first step (EBA (2023)) is to ask financial institutions to find an appropriate balance between model performance and explainability results by avoiding unnecessary complexity. This could be done, for example, by requiring developers to justify the selection of a complex model over a simpler one. This could involve showing the extent of improvement in model performance in relation to challenger models in a variety of scenarios entailing substantially different model inputs.

A more consequential decision to acknowledge that trade-off would be to allow the use of complex models that do not fully meet the established explainability standards but whose performance is unambiguously and significantly better than that of more traditional and simpler models. Arguably, prohibiting the use of such models could restrict banks and insurers from leveraging advanced technologies to manage risks effectively and improve client experience, thereby supporting socially valuable objectives such as consumer satisfaction and financial stability.

However, the introduction of explainability waivers should only affect AI models whose explainability gap is limited, with consideration for the level of riskiness of the use of such models. Moreover, it should imply the application of sufficient safeguards. For instance, frequent stability/repeatability testing,[50] also by third parties, and ongoing monitoring of model outputs could be required to ascertain whether model outputs remain consistent under varying conditions. Moreover, alternative enhanced risk management measures, data governance and human oversight may be required to compensate for insufficient explainability.[51] Firms can disclose how they adjust complex AI models to make model outputs more reliable.[52] Additionally, circuit breakers can act as automated mechanisms to halt model use in extreme or unexpected scenarios, providing a layer of protection against adverse outcomes. For use cases with significant potential impact on the safety and soundness of the institution or on consumer interests, MRM guidelines could require banks to be ready to rapidly cease the use of those models as soon as relevant performance flaws are identified.

Addressing the low explainability of AI models used for regulatory purposes is trickier. As discussed above, complex AI models may fail to meet existing MRM guidelines, including those for the use of models for regulatory purposes. As a result, the use of AI models for such purposes could be disallowed. However, this removes incentives for financial institutions to use potentially good performing models for risk management. This could hamper the development of such AI use cases, thereby sacrificing potential benefits. A compromise might be that the use of well performing complex AI models for the calculation of provisions, minimum capital or other regulatory obligations could be allowed to a limited degree. For example, it could be that such models will only be allowed for certain risk categories and exposures, or that the risk weights calculated using such models will be subject to a more stringent output floor than that currently envisaged in Basel III for more traditional internal models.

[49] OSFI and Global Risk Institute (2023) offers an alternative view that complex models may not necessarily always yield more accurate predictive performance than interpretable models.

[50] See IMDA and PDPC (2020).

[51] See IAIS (2025).

[52] In the case of LLMs, such adjustments include fine-tuning pre-trained models with data that are more relevant to the use case in question or using retrieval-augmented generation to feed them relevant document/information in order to confine the output to that specific set of information.

Section 5 – Conclusion

The use of AI is expected to become more prevalent across the business activities of financial institutions as they seek to optimise the benefits of the new technology. This means that AI applications will not just be limited to internal productivity-enhancing purposes, but could also spread to critical business areas of financial institutions. Financial authorities therefore need to be attuned to the potential implications of the use of AI for both the risks facing financial institutions individually and the risks facing the entire financial system.

The lack of explainability of certain AI models is a key concern for financial authorities. While there are several explainability techniques available to financial institutions, these have limitations when applied to more complex AI models such as LLMs. As financial institutions roll out more complex AI models in their critical business areas, this will have implications for consumers, regulatory compliance and systemic risk. In essence, this will heighten financial institutions' model risks. Specific attention is needed for closed source, proprietary models, including many LLMs for which the foundation models are not transparent. Firms using LLMs do not know how the foundation models have been trained, and this lack of explainability may limit use cases to low-risk activities.

It is therefore imperative that financial authorities seek to foster sound MRM practices in financial institutions that take into account AI developments. Authorities can do this through the issuance of MRM guidelines that:

(1) address the use of models more broadly in financial institutions, ie not just for regulatory purposes;

(2) recognise that models used or that will be used by financial institutions have evolved to include those that use AI; and

(3) reflect industry practices and adapt as these evolve.

In the context of AI explainability, for example, MRM guidelines could include requiring financial institutions to employ explainability techniques to explain black box models, establishing explainability standards based on the potential impact and riskiness of the model, and requiring complementary safeguards such as enhanced data governance and human oversight to mitigate risks associated with the use of complex AI models in critical business areas. Ultimately, there may be a need to recognise trade-offs between explainability and model performance, so long as risks are properly assessed and effectively managed. Some conditional and constrained flexibility in the application of explainability requirements for models with robustly proven good performance could be envisaged.

Authorities will also need to upskill their staff to be able to understand explainability submissions by firms. This is no trivial task, as it may be challenging to understand even inherently interpretable models, let alone the explainability techniques of black box models.

References

Alonso-Robisco, A and J Carbó (2025): "Should we trust the credit decisions provided by machine learning models?", *Computational Economics*, January.

Alonso-Robisco, A, J Carbó, G de Haro and J Guillén García (2025): "Constrained machine learning models for credit default prediction: who wins, and who loses", Florence School of Banking and Finance, *Banking Supervision Policy Working Paper Series*, no 2025/05, June.

Alvarez-Melis, D and T Jaakkola (2018): *"*On the robustness of interpretability methods", June.

Ameisen, E, J Lindsey, A Pearce, W Gurnee, N Turner, B Chen, C Citro, D Abrahams, S Carter, B Hosmer, J Marcus, M Sklar, A Templeton, T Bricken, C McDougall, H Cunningham, T Henighan, A Jermyn, A Jones, A Persic, Z Qi, T Thompson, S Zimmerman, K Rivoire, T Conerly, C Olah and J Batson (2025): "Circuit tracing: revealing computational graphs in language models", *Transformer Circuits Thread*, March.

Apley, D and J Zhu (2020): "Visualising the effects of predictor variables in black box supervised learning models", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol 82, no 4, September, pp 1059–86.

Bank of England (2025): *Financial stability in focus: artificial intelligence in the financial system*, April.

Bank of England and Financial Conduct Authority (2024): *Artificial intelligence in UK financial services – 2024*, November.

Basel Committee on Banking Supervision (BCBS) (2001): *The internal ratings-based approach – consultative document*, January.

——— (2013): *Principles for effective risk data aggregation and risk reporting*, January.

——— (2017): *High-level summary of Basel III reforms*, December.

——— (2018): *Stress testing principles*, October.

——— (2022): "CRE 36 – IRB approach: minimum requirements to use IRB approach", *Basel Framework*, December.

——— (2024a): *Core Principles for effective banking supervision*, April.

——— (2024b): "MAR 30 – Internal models approach: general provisions", *Basel Framework*, July.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency (2011): *Supervisory guidance on model risk management*, April.

Bordt, S, M Finck, E Raidl and U von Luxburg (2022): "Post-hoc explanations fail to achieve their purpose in adversarial contexts", May.

Buckmann, M and A Joseph (2023): "An interpretable machine learning workflow with an application to economic forecasting", *International Journal of Central Banking*, vol 19, no 4, October, pp 449–522.

Candelon, F, T Evgeniou and D Martens (2023): "AI can be both accurate and transparent", *Harvard Business Review*, May.

Central Bank of the United Arab Emirates (CBUAE) (2022): *Model management standards*, November.

Crisanto, J, C Leuterio, J Prenio and J Yong (2024): "Regulating AI in the financial sector: recent developments and main challenges", *FSI Insights on policy implementation,* no 63, December.

Dandl, S, C Molnar, M Binder and B Bischl (2020): "Multi-objective counterfactual explanations", in *Parallel Problem Solving from Nature – PPSN XVI*, proceedings of the 16th International Conference, PPSN 2020, Leiden, August, pp 448–69.

Daníelsson, J, R Macrae and A Uthemann (2022): "Artificial intelligence and systemic risk", *Journal of Banking & Finance*, vol 140, July.

Davis, R, A Lo, S Mishra, A Nourian, M Singh, N Wu and R Zhang (2022): "Explainable machine learning models of consumer credit risk", *The Journal of Financial Data Science*, vol 5, no 4, January, pp 9–39.

De Nederlandsche Bank and Dutch Authority for the Financial Markets (2024): *The impact of AI on the financial sector and supervision*, June.

Dhurandhar, A, K Ramamurthy, K Ahuja and V Arya (2023): *Locally invariant explanations: towards stable and unidirectional explanations through local invariant learning*, September.

Doshi-Velez, F and B Kim (2017): "Towards a rigorous science of interpretable machine learning", March.

European Banking Authority (EBA) (2023): *Machine learning for IRB models: follow-up report from the consultation on the discussion paper on machine learning for IRB models*, August.

European Data Protection Supervisor (2023): "Explainable artificial intelligence", *TechDispatch*, no 2/2023, November.

European Insurance and Occupational Pensions Authority (EIOPA) (2025): *Consultation paper on opinion on artificial intelligence governance and risk management*, February.

Federal Office for Information Security (2024): "Transparency of AI systems", white paper, August.

Financial Reporting Council (2024): *Technical actuarial guidance – models*, October.

Financial Services Agency of Japan (FSA) (2021): *Principles for model risk management*, November.

Financial Stability Board (FSB) (2024): *The financial stability implications of artificial intelligence*, November.

Gilpin, L, D Bau, B Yuan, A Bajwa, M Specter and L Kagal (2019): "Explaining explanations: an overview of interpretability of machine learning", February.

Goldstein, A, A Kapelner, J Bleich and E Pitkin (2015): "Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation", *Journal of Computational and Graphical Statistics*, vol 24, no 1, March, pp 44–65.

Heaven, W (2023): "GPT-4 is bigger and better than ChatGPT – but OpenAI won't say why", *MIT Technology Review,* March.

Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission Singapore (PDPC) (2020): *Model artificial intelligence governance framework – Second edition*, January.

Institute of International Finance and EY (IIF-EY) (2025): *IIF-EY annual survey report on AI/ML use in financial services – Public summary*, January.

International Association of Insurance Supervisors (IAIS) (2024): *Insurance Core Principles and Common Framework for the Supervision of Internationally Active Insurance Groups*, December.

——— (2025): *Application Paper on the supervision of artificial intelligence*, July.

Lakkaraju, H and O Bastani (2020): "'How do I fool you?': Manipulating user trust via misleading black box explanations", in *AIES '20*, proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, February, pp 79–85.

Lipton, Z (2017): "The mythos of model interpretability", March.

Lundberg, S and S-I Lee (2017): "A unified approach to interpreting model predictions", November.

Molnar, C (2020): *Interpretable machine learning: A guide for making black box models explainable*.

Molnar, C, G König, J Herbinger, T Freiesleben, S Dandl, C Scholbeck, G Casalicchio, M Grosse-Wentrup and B Bischl (2020): "Pitfalls to avoid when interpreting machine learning models", July.

Monetary Authority of Singapore (MAS) (2024): *Artificial intelligence model risk management – observations from a thematic review*, December.

National Institute of Standards and Technology (NIST) (2023): *Artificial intelligence risk management framework (AI RMF 1.0)*, January.

Office of the Comptroller of the Currency (OCC) (2021): "Model risk management", *Comptroller's Handbook – Safety and Soundness*, August.

Office of the Comptroller of the Currency, Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, Bureau of Consumer Financial Protection and National Credit Union Administration (2021): "Request for information and comment on financial institutions' use of artificial intelligence, including machine learning", *Federal Register*, vol 86, no 60, March.

Office of the Superintendent of Financial Institutions (OSFI) (2023): *Draft guideline E-23 – Model risk management*, November.

OSFI and Global Risk Institute (2023): *Financial industry forum on artificial intelligence: a Canadian perspective on responsible AI*, April.

Pearl, J and D Mackenzie (2018): *The book of why: the new science of cause and effect*, May.

Prenio, J and J Yong (2021): "Humans keeping AI in check – emerging regulatory expectations in the financial sector", *FSI Insights on policy implementation*, no 35, August.

Prudential Regulation Authority (PRA) (2023): "Model risk management principles for banks", *Supervisory Statement*, no SS1/23, May.

Retzlaff, C, A Angerschmid, A Saranti, D Schneeberger, R Röttger, H Müller and A Holzinger (2024): "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists", *Cognitive Systems Research*, vol 86, August.

Ribeiro, M, S Singh and C Guestrin (2016): "'Why should I trust you?' Explaining the predictions of any classifier", August.

——— (2018): "Anchors: high-precision model-agnostic explanations", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol 32, no 1, April.

Rudin, C (2019): "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, vol 1, May, pp 206–15.

Russell, C, S Wachter and B Mittelstadt (2018): "Counterfactual explanations without opening the black box: automated decisions and the GDPR", March.

Saleem, R, B Yuan, F Kurugollu, A Anjum and L Liu (2022): "Explaining deep neural networks: A survey on the global interpretation methods", *Neurocomputing*, vol 513, November, pp 165–80.

Swiss Financial Market Supervisory Authority (FINMA) (2024): *FINMA Guidance 08/2024 – Governance and risk management when using artificial intelligence*, December.

Thampi, A (2022): *Interpretable AI: Building explainable machine learning systems*, July.

Thomas, R (2024): "Unmasking generative AI: Understanding explainability techniques", August.

Wei, J, X Wang, D Schuurmans, M Bosma, B Ichter, F Xia, E Chi, Q Le and D Zhou (2023): "Chain-of-thought prompting elicits reasoning in large language models", January.

Wu, Y, M Keoliya, K Chen, N Velingker, Z Li, E Getzen, Q Long, M Naik, R Parikh and E Wong (2024): "DISCRET: Synthesizing faithful explanations for treatment effect estimation", June.