Google Cloud

air Alliance for Innovative Regulation

Applying model risk management guidance to artificial intelligence/ machine learning-based risk models

Updated: June 2023

1. Executive summary

Advances in artificial intelligence (AI) and machine learning (ML) have led to increased adoption in the financial services sector. A prominent use for this technology is to assist in key compliance and risk functions, including the detection of fraud, money laundering, and other financial crimes, as well as trade manipulation—collectively raeferred to as "Risk AI/ML."[1] As the use of these models grows, so do questions about managing risks associated with the models. In particular, regulators, financial institutions, and technology service providers have been looking at whether existing model risk management guidance[2] ("MRM Guidance")—which has traditionally been the regulatory regime applicable to managing model risk in the financial services industry—continues to be relevant for AI/ML models and, if so, how the guidance should be interpreted and applied to this new technology.

This white paper seeks to address that question, with the aim of fostering thought and dialogue among agencies, financial institutions, risk model vendors, and other entities interested in the performance, outputs, and compliance of models used to identify, mitigate, and combat risks in the financial services industry. However, it does not purport to address specific issues that may arise with other applications of AI/ML, such as consumer credit underwriting, or models incorporating the recent advances in generative AI technology.

At the outset, the paper argues that MRM Guidance, given its broad, principles-based approach, continues to provide an appropriate framework for assessing financial institutions' management of Risk AI/ML models. Nonetheless, this white paper recognizes that AI/ML models have some unique traits and characteristics compared to conventional models, including their potential dynamism and pattern recognition capabilities. These distinctions must be in focus when considering how MRM Guidance should be applied to Risk AI/ML models.

Taking into account those unique aspects of AI/ML models, this paper offers specific observations and recommendations regarding the application of MRM Guidance to Risk AI/ML models, including:

- **Risk assessment**
  In assessing the risk presented by a model, it is important to recognize that all AI/ML models are not inherently more risky than conventional models. A risk-tiering assessment must consider the targeted business application or process for which a model is used, as well as the model's complexity and materiality. To assist in these assessments, regulators could clarify that the use of AI/ML alone does not place a model into a high-risk tier and publish further guidance to help set expectations regarding the materiality/risk ratings of AI/ML models as applied to common use cases.

[1] These use cases are distinguished from other applications that may raise unique policy and regulatory issues, including AI/ML models used in the context of consumer finance underwriting. They are also distinguishable from models using generative AI or Large Language Models, also known as Foundational Models. The question of appropriate risk management approaches for those categories of AI/ML models are beyond the scope of this paper, but may be addressed iteratively through subsequent efforts and building on some of the principles discussed herein.

[2] For example, in the United States, OCC, Bull. 2011-12 (April 4, 2011), available at https://occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html; and OCC, Bull. 2021-39 (Aug. 18, 2021), available at https://occ.gov/news-issuances/bulletins/2021/bulletin-2021-39.html.

- **Safety and soundness**
  Due to the dynamic nature of Risk AI/ML models, reliance on extensive and ongoing testing focused on outcomes throughout the development and implementation stages of such models should be primary in satisfying regulatory expectations of fitness and soundness. To that end, the development of technical metrics and related testing benchmarks should be encouraged. Model "explainability," while useful for purposes of understanding specific outputs of AI/ML models, may be less effective or insufficient for establishing whether the model as a whole is sound and fit for purpose.

- **Model documentation**
  The touchstone for the sufficiency of model documentation should be what is needed for the bank to use and validate the model, and to understand its design, theory, and logic. Disclosure of proprietary details, such as model code, is unnecessary and unhelpful in verifying the sufficiency of a model and would deter model builders from sharing best-in-class technology with financial institutions.

- **Industry standards and best practices**
  Regulators should support the development of global standards and their use across the financial services and regulatory landscape by explicitly recognizing such standards as presumptive evidence of compliance with the MRM Guidance and sound AI/ML risk mitigation practices. In addition, regulators should foster industry collaboration and training based on such standards.

- **Governance controls**
  Regulators should use guidance to advance the use of governance controls, including incremental rollouts and circuit breakers, as essential tools in mitigating risks associated with Risk AI/ML models.

2. Introduction

Catalyzed by the global pandemic, businesses are accelerating their digital transformation, including through increased adoption of AI/ML-based technologies. According to the International Data Corporation (IDC), global spending on AI systems is expected to hit $154 billion in 2023 and is forecasted to increase to more than $300 billion in 2026, with a compound annual growth rate (CAGR) of 27% over the period.[3] The banking industry is expected to be one of the top industries investing in this technology, particularly in applications designed to reduce risk and support regulatory compliance, such as automated threat intelligence and fraud analysis applications. Indeed, financial regulators themselves are increasingly looking to adopt new AI/ML technologies to promote efficient and effective oversight, including through analysis of increasing volumes of data.[4] In the United States, for example, the Financial Industry Regulatory Authority (FINRA) recently unveiled a comprehensive, new market surveillance program leveraging advanced AI technologies to ensure market integrity.[5]

The adoption of AI/ML technologies can provide significant benefits. As federal banking agencies in the United States acknowledged in a recent Request for Information (RFI):

> AI has the potential to offer improved efficiency, enhanced performance, and cost reduction for financial institutions, as well as benefits to consumers and businesses. AI can identify relationships among variables that are not intuitive or not revealed by more traditional techniques. AI can better process certain forms of information, such as text, that may be impractical or difficult to process using traditional techniques. AI also facilitates processing significantly large and detailed datasets, both structured and unstructured, by identifying patterns or correlations that would be impracticable to ascertain otherwise. AI applications may also enhance an institution's ability to provide products and services with greater customization.[6]

As a result, the range of AI/ML use cases in financial services is broad and growing. For example, banks are increasingly looking to leverage AI/ML technology to improve costs and user experiences in their call centers. According to industry analysis, in 2023, 40% of contact center interactions will be fully automated by using such AI capabilities as personalization, AI-based customer routing, language sentiment analysis, intelligent document processing, workforce management, post call wrap-up, and task and process workflow automation.[7] In other cases, AI/ML is being used to automate data capture from documents, significantly streamlining and speeding up processes such as mortgage lending. As another example, a recent industry survey of fraud and compliance professionals noted that "while only 17% of organizations' anti-fraud programs currently use artificial intelligence or machine learning analytics, these techniques are expected to experience [significant] growth, with 26% anticipating that their organizations will adopt this type of advanced analytics technology in the next two years."[8] This pool of use cases can be expected to further expand in the future as firms look for ways to process, analyze, and generate actionable insights from large bodies of data.

[3] IDC, Worldwide Spending on AI-Centric Systems Forecast to Reach $154 Billion in 2023 (Mar. 7, 2023), available at https://www.idc.com/getdoc.jsp?containerId=prUS50454123#:~:text=Worldwide%20Spending%20on%20AI%2DCentric,in%202023%2C%20According%20to%20IDC.

[4] See Jo Ann Barefoot, The case for placing AI at the heart of digitally robust financial regulation (May 24, 2022), Brookings, available at https://www.brookings.edu/research/the-case-for-placing-ai-at-the-heart-of-digitally-robust-financial-regulation/.

[5] See FINRA, Deep Learning: The Future of the Market Manipulation Surveillance Program (podcast) (Jan. 25, 2022), available at https://www.finra.org/media-center/finra-unscripted/deep-learning-market-surveillance.

[6] Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 86 Fed. Reg. 16,837 (Mar. 31, 2021).

[7] Gartner: Forecast Analysis: Contact Centers, Worldwide (April 26, 2019) and Gartner: Gartner Identifies Three Important Ways AI Can Benefit Customer Service Operations (Jan 19, 2022).

[8] 2022 Anti-Fraud Technology Benchmarking Report, Association of Certified Fraud Examiners (ACFE).

No technology is without risk, however. And, as use cases grow in financial services, so does the need to assess and mitigate risks associated with the use of AI/ML-based tools, some of which are unique to AI/ML while others are not. Financial regulatory agencies in most countries have not issued rules or guidance specific to AI/ML, but rather, have sought to apply existing "model risk management" guidance and principles to the technology.

In the United States, for example, the "Supervisory Guidance on Model Risk Management" is relevant and applied with respect to AI/ML-based models. This guidance was issued jointly by the Federal Reserve Board of Governors and the Office of the Comptroller of the Currency ("OCC") in 2011, in the wake of the global financial crisis, and before AI/ML-based models were in significant use.[9] While more recent supplemental guidance has been issued that touches on AI/ML in some jurisdictions,[10] including in the form of the "Model Risk Management" Handbook released by the OCC in 2021,[11] many stakeholders have asked whether current regulatory guidance in the financial services industry is sufficient for addressing risk arising from the use of AI/ML-based models.[12] A number of countries, including the United States, Germany, and Canada, have been exploring this question through regulatory comment processes.[13]

This white paper explores the application of the U.S. model risk management guidance ("MRM Guidance"), including both the 2011 and 2021 releases, to AI/ML-based models deployed in the banking industry. Its purpose is to offer input to the financial regulatory authorities charged with issuing and updating MRM guidance in today's dynamic technology environment. We also hope to foster thought and dialogue among agencies, the banking industry, risk model vendors, and other entities interested in the performance, outputs, and compliance of models used in finance for purposes such as countering fraud and illicit activity.

With respect to AI/ML model use cases, we focus specifically on risk mitigation and regulatory compliance applications (referred to herein as "Risk AI/ML"), such as models designed to detect fraud and money laundering. We apply this focus because Risk AI/ML models are among the most common AI/ML models under development. Further, AI/ML models used for other applications, such as credit risk assessment, require grappling with specialized policy and regulatory issues relating to consumer finance and fair lending requirements, which add further complexity. Those categories of models, as well as models using generative AI or Large Language Models (also known as Foundational Models), are beyond the scope of this white paper. The question of appropriate risk management approaches for those categories of AI/ML models should be addressed iteratively through subsequent efforts and building on some of the principles discussed here.

As a threshold matter, this white paper takes the view that the MRM Guidance provides an appropriate framework for assessing and mitigating risk associated with Risk AI/ML models.[14] The guidance is sufficiently principles-based to continue to be valid and useful even in this new context. Wholesale additional rulemakings specific to Risk AI/ML are not needed and would likely generate significant new questions and uncertainties. Rather, we believe financial regulators should use existing authority and regulatory tools (e.g., supervisory guidance and

[9] OCC, Bull. 2011-12 (April 4, 2011), available at https://occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html.

[10] We note that the U.S. Department of Commerce's National Institute of Standards and Technology (NIST) recently released its Artificial Intelligence Risk Management Framework (AI RMF 1.0), which is focused on helping organizations manage AI risks. For purposes of this white paper, we will focus on sector-specific regulation, but applaud and support horizontal and sector-agnostic efforts, including the AI RMF 1.0.

[11] For purposes of this white paper, we will refer to the 2011 Guidance and the 2021 Handbook collectively as the "MRM Guidance."

[12] OCC, Bull. 2021-39 (Aug. 18, 2021), available at https://occ.gov/news-issuances/bulletins/2021/bulletin-2021-39.html.

[13] See AI Risk Management Framework | NIST, National Institute of Standards and Technology, available at https://www.nist.gov/itl/ai-risk-management-framework; DE.DIGITAL - Federal Government adopts Artificial Intelligence Strategy; German Artificial Intelligence Strategy, available at https://www.de.digital/DIGITAL/Redaktion/EN/Meldungen/2018/2018-11-16-federal;-government-adopts-artificial-intelligence-strategy.html; Pan-Canadian AI Strategy – CIFAR, Pan-Canadian Artificial Intelligence Strategy, available at https://cifar.ca/ai.

[14] We note that the MRM Guidance applies, expressly, to a subsection of the financial services industry. However, even in other contexts in which it does not expressly apply, it may serve as a useful touchstone in the development of model governance approaches.

alerts, training, and robust industry engagement) to clarify how technology providers, banking institutions, and banking examiners should apply the existing MRM Guidance to ensure proper risk management of Risk AI/ML-based models.

To that end, this white paper is structured to: (1) outline how AI/ML models differ from more conventional models used in banking, (2) discuss how elements of the MRM Guidance apply to Risk AI/ML models, and (3) offer observations and recommendations that can clarify and validate risk mitigation practices pertaining to Risk AI/ML models consistent with the MRM Guidance.

3. What is an AI/ML-based model and how is it different?

AI is a convenient informal term, but has no universally agreed-upon technical meaning.[15] Most modern AI applications are, however, based on ML, which does have a well-accepted technical definition—namely, techniques that let software learn from example data, rather than from rules defined by programmers. Put simply, ML is a new way of creating problem-solving systems. The goal is to create a model, or a mathematical representation of patterns, that adjusts over time as more examples are examined and more is learned about the patterns. As a result, a key characteristic of AI/ML-based models is that they are potentially dynamic[16] and capable of adapting to data inputs over time. This is also a key differentiation between AI/ML-based models and many conventional statistical models.

Other key characteristics of AI/ML-based models include their high dependency on quality data and their ability to identify correlations among and across large data sets. Because machine learning helps identify patterns in data, which are then used to make predictions about new data points, proper construction and sufficiency of data sets (examples) for training the model are particularly critical. The ability to consume and analyze large datasets holds substantial promise with respect to regulatory compliance, oversight, and transaction monitoring applications.[17] For example, financial authorities invest significant effort in searching for patterns that may reflect risk trends or noncompliance, including those related to illegal trading activity, money laundering, and fraud.

In addition, how an AI/ML-based model reaches a decision or generates an output may be more complex than that of more conventional models. A rules-based model designed to filter large volumes of transactions to identify potential money laundering, for example, will associate transactions to defined rules that are programmed into the model (e.g., flag when multiple funds transfers are sent in large, round dollar, hundred dollar, or thousand dollar amounts).

By contrast, AI/ML is often a productive solution for business processes that rely on pattern recognition, which may correlate to known patterns but may also require identification of new or emerging patterns. This characteristic is especially important with Risk AI/ML models given the adversarial contexts in which they are often deployed. In the anti-money laundering scenario noted above, for example, a rules-based model could simply incentivize bad actors to shift behaviors to avoid a flag. In contrast, an AI/ML model—which may better enable the detection of

[15] See Council of Europe, What's AI?, available at https://www.coe.int/en/web/artificial-intelligence/what-is-ai.

[16] Even these models, though, often require human involvement to achieve such dynamism.

[17] See Jo Ann Barefoot, The case for placing AI at the heart of digitally robust financial regulation (May 24, 2022), Brookings, available at https://www.brookings.edu/research/the-case-for-placing-ai-at-the-heart-of-digitally-robust-financial-regulation/.

new illicit patterns as well as known typologies—may be more difficult for bad actors to evade.

Finally, while perhaps not an intrinsic characteristic of AI/ML models, it is nonetheless common that AI/ML models in the banking sector are developed through collaborations between banking institutions and technology providers. Such collaborations allow the banking system to incorporate leading AI/ML technologies. But they also implicate aspects of the MRM Guidance related to third-party vendors, requiring banks, vendors, and examiners to address an additional set of issues, such as how contractual and compliance responsibilities and ongoing product maintenance requirements should be distributed between banks and vendors. It also requires assessment of the degrees of, and approaches to, information-sharing that will be required to satisfy MRM requirements.

These overall characteristics of AI/ML technologies have implications for the application of MRM Guidance to Risk AI/ML models. As detailed further below, AI/ML technologies used in risk models may require more emphasis on—and prioritization of—certain model validation and risk mitigation practices to establish their soundness (e.g., robust, independent, and ongoing testing) as compared to conventional rules-based models. To this end, regulatory recognition of appropriate deployments of certain model risk management tools in the AI/ML context would have benefits both for the industry and for regulators, including providing greater certainty as to how regulators are evaluating the risks presented by AI/ML applications and related mitigation techniques. This clarity would subsequently encourage greater adoption and use of these applications in regulated industries.

> **Takeaway**
>
> AI/ML models have some unique traits and characteristics as compared to conventional statistical models, including their potential dynamism and pattern recognition capabilities. These distinctions need to be in focus when considering how MRM Guidance should be applied to Risk AI/ML models. Certain aspects of the MRM Guidance may take on greater importance relative to Risk AI/ML models, including an emphasis on robust, diverse, and ongoing testing.

4. Model risk management guidance requirements

As noted above, the MRM Guidance issued by the U.S. banking regulatory agencies is the relevant framework for supervised entities to promote model compliance and reduce model risk in the U.S. financial services industry. The MRM Guidance establishes 11 criteria relating to model development: (1) Board and Senior Management Oversight, (2) Personnel, (3) Policies and Procedures, (4) Planning, (5) Assessing Risk, (6) Model Inventory, (7) Documentation, (8) Data Management, (9) Model Development, Implementation, and Validation, (10) Third-Party Risk Management, and (11) Internal Audit.

Risks associated with the development and use of an AI/ML model, although varied depending upon the specific model and its application, are identified and managed in many respects using the same control processes employed for more conventional models at each stage of the model development lifecycle (i.e., development, implementation, use, validation). Thus, for many of these requirements, a different approach

may not be required as between a conventional model and an AI/ML-based model.

However, given the unique characteristics of AI/ML-based models discussed in the section above, how some aspects of the MRM Guidance should apply to AI/ML models may be less clear. And in other cases, these unique characteristics may require unique approaches. Additional clarity may be needed in such cases, which is hardly unexpected given that the MRM Guidance is intentionally broad. Nevertheless, this breadth can also lead to disparate application of the Guidance, especially in the context of novel AI/ML models, and can produce uneven expectations depending on the particular focus of an examiner or compliance team.

With respect to conventional models, stakeholders have had more than a decade to confirm how best to consider and apply the MRM Guidance in order to mitigate particular model risks. Given the relative recency of AI/ML models, however, an opportunity exists to similarly confirm and tailor the MRM Guidance to the unique attributes and characteristics of AI/ML models. The following observations and recommendations are intended to generate additional clarity regarding the application of select MRM Guidance requirements to Risk AI/ML models to promote operational compliance and regulatory certainty.

5. Observations and recommendations for applying MRM guidance to risk AI/ML models

A. Use of AI/ML technology by itself should not render a model "high risk"

A foundational requirement of the MRM Guidance is that the risk associated with each model must be assessed and models placed into distinct risk categories or tiers so that model risk management resources and mitigation techniques can be properly allocated. The MRM Guidance states that this same approach applies with respect to AI/ML-based models. Specifically, according to the MRM Guidance, "[r]isk management of AI, as with any other innovative technology, should be commensurate with the materiality and complexity of the model or tool and the activity's risk or business process that the AI is supporting."[18] Put simply, risk assessment determinations are critical in shaping the direction and intensity of compliance efforts.

This is an important regulatory assertion and underlying it should be clear recognition that the mere fact that a model utilizes AI/ML technologies does not necessarily make it risky. An AI/ML model used to supplement or replace a legacy model may pose no increased risk, for example, if it matches the performance of the legacy model through robust parallel testing and rollout, and may even incrementally improve performance. Such an acknowledgment may be particularly important to counter perceptions that AI/ML technology is inherently complex and therefore presents significant risk.

[18] OCC Model Risk Management (2021), at 13.

Additionally, beyond complexity and materiality, the Handbook underscores that it is critical to consider the activity and/or business process of which the AI/ML-based model is part in order to help calibrate the overall level of risk. Even if an AI/ML model were complex, considerations of its application may predominate in assessing overall risk.

To illustrate, consider the example of a model that utilizes AI/ML for the purposes of textual analysis, such as a document parser. Such a tool may help pull out important information from a mortgage loan application, for example, to help bank officials more quickly process and act upon mortgage loan requests. In this example, the language-parsing model may certainly be complex—such models often utilize computer vision and natural language processing and may be trained on extensive volumes of relevant documents across lending, insurance, government, and other industries. However, given their limited function (they assist in document processing but not decisioning) and the fact that they are almost always part of a broader process with humans in the loop, it would be difficult to categorize such a model as high- or even medium-risk.

The overall risk profile of this "Doc-AI" model would likely be different from that of an AI/ML-based fraud or anti-money laundering model designed to identify fraudulent or illegal behavior, and different still from an AI/ML-based model used to support credit decisioning. Every one of those models is likely complex, but the assessment of its overall risk categorization needs to account for the additional factors described in the guidance and properly consider the totality of the factors.

With respect to materiality of general categories of model application and use cases, banking authorities might consider clarifying or confirming industry risk-tiering considerations. To this end, further regulatory guidance might include the explicit recognition of "targeted use" classifications for models (e.g., risk management, consumer services, pricing and valuation practices, fraud and transaction monitoring, marketing activities, others) and the regulatory model "materiality" ratings (e.g., low, medium or high risk) typically assigned. These assignments and materiality determinations should be supported by clear regulatory criteria to help stakeholders anticipate regulatory expectations. Guidance could further highlight potential risks commonly associated with such targeted use cases. This approach would assist in aligning the interests of AI/ML model stakeholders regarding the identification and materiality, prioritization, and mitigation of such model risks.

Takeaway

AI/ML models are not inherently more risky than conventional models. A risk-tiering assessment must consider the targeted business application or process for which a model is used, as well as model complexity and materiality. To assist in these assessments, regulators could clarify that the use of AI/ML alone does not place a model into a high-risk tier and publish further guidance to help set expectations regarding the materiality/risk ratings of AI/ML models as applied to common use cases.

B. Robust testing of AI/ML models takes on increased importance relative to explainability in establishing soundness and fitness for purpose

The MRM Guidance suggests a number of factors to be considered in assessing a model's soundness and fitness for purpose.[19] To this end, the MRM Guidance frequently references "explainability" as one of the key considerations with respect to AI/ML models. For example, the MRM Guidance provides that, with respect to the requirement that financial institutions have sufficient risk assessment protocols, it is important for bank examiners to "assess if [AI/ML-based] model ratings take explainability into account."[20] Similarly, with respect to the requirement that financial institutions evaluate the "conceptual soundness" of AI/ML-based models, the MRM Guidance provides that:

> Transparency and explainability are key considerations that are typically evaluated as part of effective risk management regarding the use of complex models. The appropriate level of explainability of a model outcome depends on the specific use and level of risk associated with that use. ... There may be challenges with explaining some models based on complexity or, in some cases, limited documentation provided for third-party models. Examiners should discuss with bank management the bank's process for exploring various approaches to determine whether bank personnel have an understanding of how models function and make decisions, including identifying any limitations and use of compensating controls.[21]

The MRM Guidance goes on to define "explainability" as "the extent to which AI decisioning processes and outcomes are reasonably understood by bank personnel."

Explainability, as defined in this way, plays an important role in risk assessment. For example, a bank using an AI/ML-based model focused on identifying anomalous behavior suggesting money laundering may want to know what factors led to the flagging of a particular transaction for review. And the technology around explainability—referred to as "explainable AI" technology—is evolving significantly to meet these needs (see Deep Dive: Explainability).

While explainability is useful for the purposes of understanding specific outcomes of AI/ML models, it may be ineffective or insufficient for establishing whether the model itself is sound and fit for purpose. Specifically, unlike conventional models (e.g., linear regression models) where relatively simple explainability techniques can both help to demonstrate how the model works and how specific outcomes were determined (e.g., because of a reliance on if/then rules or decision trees), more complex AI/ML models often rely on explainability techniques that are able to reveal information about a particular outcome or prediction, but not necessarily whether the model is performing as it should.

Consider the example of models designed to detect money laundering. A rules-based model would be used to identify transactions that meet certain predefined criteria. Tracing specific outputs (flagged transactions) to specific inputs (identified rules) could be a strong indicator of sufficiency and soundness given the type

[19] OCC Model Risk Management (2021), at 39-40.
[20] Id. at 24.
[21] Id. at 40.

of model involved. For this instance, explaining the model's if/then logic can directly help in understanding model outcomes.

For an AI/ML-based model that looks more holistically at patterns—whether detecting anomalies or comparing patterns in a particular case to known problem cases—there are no foundational "rules" or inputs that can be specifically identified and linked to establish soundness and suitability. Indeed, this is one of the advantages of AI/ML-based approaches—that they can be scaled and adapted to new scenarios in a way that rules-based systems may not be able to do.

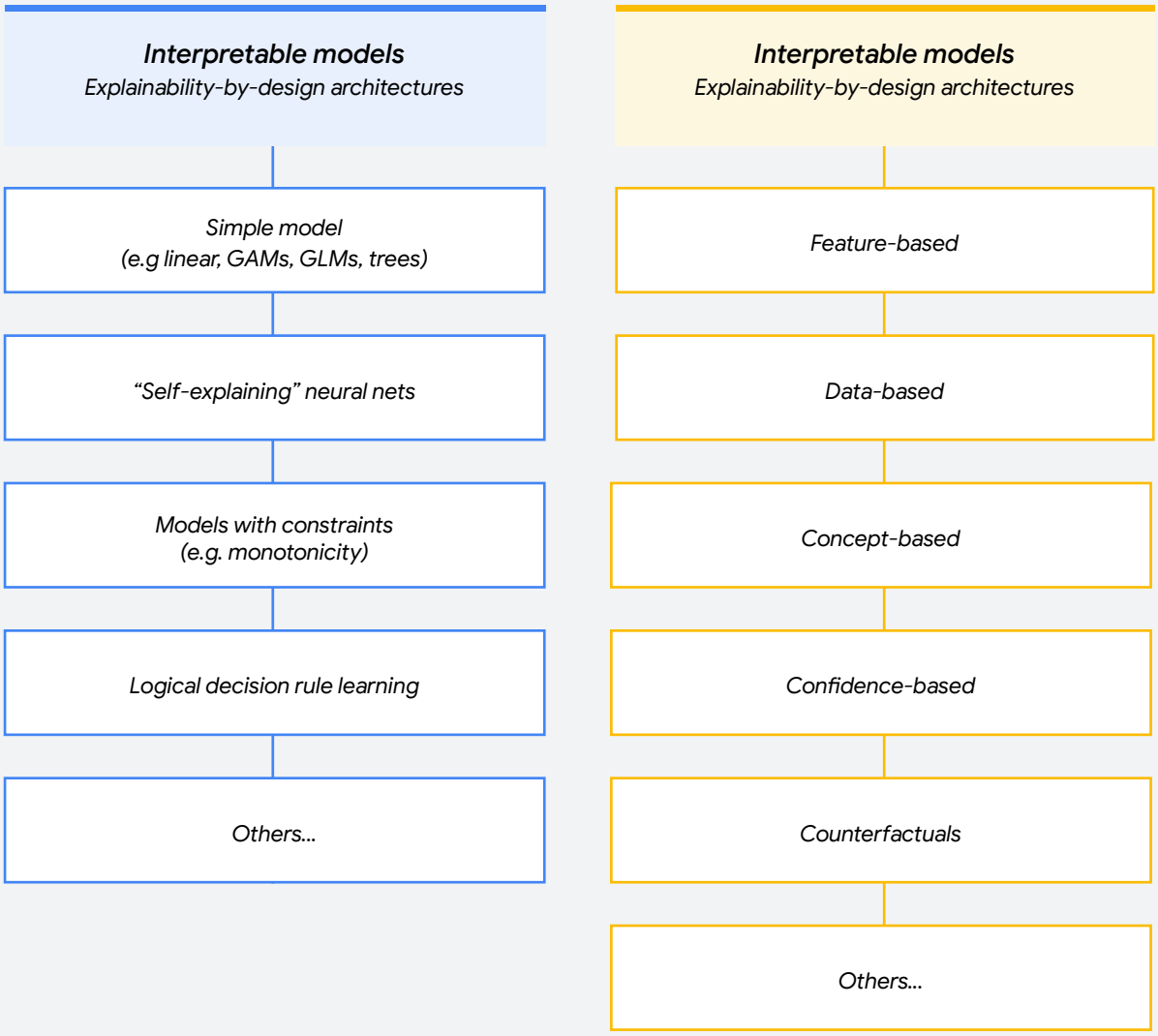
Paper continues on the next page.

Deep Dive: ***Explainability***

The term "explainability" evokes—and sometimes is used interchangeably with—a number of different but related concepts including "interpretability," "auditability," "traceability," "contestability," "accountability," and "transparency."

There are a number of technological approaches being developed to advance explainability determinations. From a technological perspective, explainability can be broadly analogized to a tree with two major branches.

AI explainability



On one branch are AI/ML models that are designed and built to be directly interpretable. This includes simple models such as linear regression models, which can be "explained" by looking at the coefficients that exactly describe the relationship between input features and model outputs. Basic decision trees also can be explained by looking at the path taken through the tree to arrive at a particular decision—the combination of rules that determine the output. Interpretable models increasingly include more flexible models that have built-in interpretability affordances, including, for example, "self explaining neural nets" that use "attention mechanisms" or constrained architectures to track what parts of a datapoint the model is keying in on most to make a particular prediction.[22]

The second branch consists of so-called "post-hoc" explainers, which involve techniques for understanding the behavior of highly complex AI/ML models. The significant predictive power of these models comes from the fact that their internal logic is not easily reduced to simple rules. For these models, many techniques are being developed to understand why a model has made certain predictions.

For example, "feature-based" explanations try to explain why a model has made a particular prediction by quantifying how much each input feature contributed to the model's prediction. For a model trained to predict the likelihood that an airline flight will be delayed, the weather is likely to be a very important input feature, whereas the average age of the passengers is likely not important. In this example, the percentage of a prediction attributable to weather might be a significant percentage contributor to the overall score.

The current state of technological development allows for the identification of methodologies that can help provide insights into the outputs of particular types of AI/ML models. Clarity from regulators regarding their explainability concerns for a particular model application and the objectives they are seeking to meet can help industry participants assess which explainability approaches and methodologies are appropriate in any given situation.

[22] Rishaba Agarwal et al., Neural Additive Models: Interpretable Machine Learning with Neural Nets (2021), available at https://proceedings.neurips.cc/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf.

Accordingly, in contrast to static rules-based models, model development and implementation for AI/ML-based models is a fluid process that includes many individual phases that must be evaluated to ensure a valid model is produced for the intended purpose and use. Testing for desired performance and outcomes of material model components at each stage of the model development lifecycle is helpful to demonstrate that individual model components, as well as the model overall, are performing as expected. Notably, as discussed further below, there are many different kinds of tests that can be deployed to ensure proper model performance and outcomes. In this way, robust, diverse, and ongoing testing of models should be a central and prioritized way to ensure AI/ML model soundness and suitability, especially relative to conventional models.

Testing for a Risk AI/ML model should be conducted at various stages of development and implementation and use various testing approaches. The MRM Guidance states, "[t]he nature of testing and analysis will depend on the type of model and will be judged by different criteria depending on the context." Additionally, "[d]ifferent tests have different strengths and weaknesses under different conditions. Any single test is rarely sufficient, so banks should apply a variety of tests to develop a sound model." Particular testing highlighted in the MRM Guidance includes checking the model's accuracy; evaluating the model's behavior over a range of input values, including extreme values; testing of judgmental or qualitative aspects of the model; and testing to take into account new data, techniques, or changes due to the deterioration of the model's performance.

In addition, the potentially more complex and dynamic nature of AI/ML models makes it particularly important to risk-assess process controls surrounding the type and ongoing use of data for such models. As a result, holistic risk management and ongoing monitoring is required to ensure that new data does not cause a model to drift or become "overfitted" (e.g., new datasets must include not just the data the model succeeds on, but additional data that is relevant to the overall theory of the model more generally, and which does not cause the model to be guided to a particular outcome). It may also be helpful to support the development of technical metrics and shared test sets that are recognized by regulators or accepted by them in order to create common benchmarks for testing. Such technical metrics and testing benchmarks could be used to assess whether there is proper alignment between model output and business goals, and also act as a control on whether data quality issues exist.[23]

When a vendor model is being considered, banks are expected to demonstrate "appropriate due diligence on the third-party relationship and the model itself."[24] In addition to the vendor/bank cooperation concepts regarding model development, the MRM Guidance also suggests that vendors provide appropriate testing results that demonstrate the model works as expected and the model performance meets the bank's needs.[25] The MRM Guidance points out that "[e]xternal models may not allow full access to computer coding and implementation details, so the bank may have to rely more on sensitivity analysis and benchmarking."[26]

Accordingly, the importance of model testing for demonstrating soundness is further heightened in the context of vendor-developed models. Naturally,

[23] An example of such technical metric standards development that can provide helpful clarity to stakeholders is NIST's Face Recognition Vendor Testing program, available at https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.

[24] OCC, Bull. 2021-39 (Aug. 18, 2021), https://occ.gov/news-issuances/bulletins/2021/bulletin-2021-39.html. The most recent OCC guidance cites back to earlier OCC Bulletins 2013-29 and 2020-10 relating to third-party risk management guidance. Id. at 48.

[25] Notably, in line with this guidance, vendors are increasingly developing AI assurance tools. See The roadmap to an effective AI assurance ecosystem (Dec. 8, 2021), available at https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem.

[26] OCC, Bull. 2021-39 (Aug. 18, 2021), at 49.

sensitivity analysis and benchmarking must be appropriate for the particular model or other types of testing, and validation checks should be employed.

As an additional consideration, and specific to third-party developed models, ongoing testing and updating practices should be carefully considered and clear responsibilities established between the model vendor and the model user. To this end, such practices should be documented and allocate responsibilities appropriate to each party. Additionally, any issues with model performance and outputs must be flagged and addressed swiftly in line with best practices.

> **Takeaway**
>
> Due to the dynamic nature of Risk AI/ML models, including those developed by third-party vendors, reliance on extensive and ongoing testing focused on outcomes throughout the development and implementation stages of such models should be recognized by banking authorities as primary in satisfying regulatory expectations of soundness. The development of technical metrics and related testing benchmarks should be encouraged, including through regulatory recognition of such standards. Additionally, testing must be robust, diverse, and validated through accepted methods.[27]

Paper continues on the next page.

[27] "There may be challenges with explaining some models based on complexity or, in some cases, limited documentation provided for third-party models. Examiners should discuss with bank management the bank's process for exploring various approaches to determine whether bank personnel have an understanding of how models function and make decisions, including identifying any limitations and use of compensating controls." Id. at 40.

Deep Dive: ***Testing protocols***

Robust and diverse testing of Risk AI/ML models should be a primary way to establish model soundness and fitness. Testing protocols can help ensure conceptual model validity, ongoing model performance monitoring, and outcome analysis. The following are examples of testing protocols that can satisfy these key objectives and be broadly deployed to ensure proper model performance and outcomes.

- ***Regression Testing***
  This testing approach uses past observed data where the risk analyst knows the outcome and is able to ensure that a new AI/ML model will yield the same or similar results. Deviation from past outcomes is not inherently improper, but should be analyzed within a framework that captures such deviations. This testing protocol covers concepts commonly referred to as "back testing," "output trajectory," "tests for overfitting," and "segmented performance" (when the "past observed data" has been categorized into human-intelligible segments).

- ***Unit Testing***
  This protocol leverages synthetic input data, often referred to as "test cases," such as hand-generated data or data hand-picked from past observations where the risk analyst knows what outcomes are expected even if she cannot articulate the exact rules. This approach also covers the concept commonly referred to as "benchmarking." If the test cases are selected such that the risk analyst believes them to cover on-the-edge cases, this protocol would also cover the concept of "above-the-line and below-the-line testing."

- ***Fuzz Testing***
  This protocol, which has similarities to unit testing, inputs intentionally invalid, unexpected, or adversarial synthetic data into the model to test for breakages, failures, or other improper outcomes. Model failure or the production of largely unexpected outcomes will trigger flags for the risk analyst and will require further assessment. This approach largely includes concepts such as "sensitivity analysis and stress testing" and "adversarial testing."

- ***Continuous Evaluation Testing***
  This approach includes the running of previous versions of AI/ML models with the most recently available data in order to flag outputs that are different from the mostly recently deployed model (i.e., the opposite of regression testing). This would cover tests for "output trajectory" and "input and prediction drift" from a different angle.

C. Model documentation requirements for third-party models should reflect pragmatic risk and intellectual property considerations

The MRM Guidance requires that an extensive set of documentation be maintained for models, but there are often significant questions about what constitutes "adequate documentation" of the model design, theory, and logic. This is especially true when it comes to documentation expectations for third-party model developers.

To this end, the Guidance expressly recognizes that in the case of models sourced from third parties, the documentation may be different than for models developed in-house. Specifically:

> When a bank uses third-party models, the extent of documentation that the bank has is typically not as extensive as for models developed in-house. Examiners should determine if documentation is sufficient for bank management to appropriately use and validate third-party models.

Limited additional guidance is provided as to the sufficiency of documentation, creating uncertainty and potential for friction between banks and technology vendors. While this friction may exist in any situation involving a third-party vendor, the potential is heightened with certain AI/ML-based models in which the particular design of the model—for example, the coding developed to train a model and feature-level data—may be critical proprietary information.[28]

Against this background, the MRM Guidance generally expects that the design, theory, and logic of the model (i.e., model methodologies, processing components, and mathematical specifications, including merits and limitations) should be "explained in detail" and, regarding published research to support the model use, that a "comparison with alternative theories and approaches is a fundamental component of a sound modeling process." How each of these regulatory requirements may be reasonably satisfied is unclear. This is especially true when it comes to published research for use cases of first impression that rely upon newer technologies and the use of AI/ML models.

Confirmation of regulatory expectations relating to documentation sufficiency would be particularly helpful to developers and users of AI/ML models. The recent OCC Guidance is a step in the right direction, but more can be done to provide certainty in the marketplace, especially in the context of third-party model developers. For example, although the model use requirement to develop and maintain "adequate documentation" that "explains in detail" the design, theory, and logic of the model is understandable and an expected industry practice for business purposes regardless of regulatory requirements, few concrete parameters are provided regarding the extent and type of documentation that would be deemed appropriate by banking examiners. Examples of sufficient documentation might therefore be indicated and useful.

The touchstone for sufficiency of documentation should be whether the documentation is required by bank management to use and validate the model. Additionally, "adequate documentation" should be measured based on the materiality of the model and the associated risks that must be considered, assessed, and mitigated related to the design, theory, and logic of the model.

[28] Former Federal Reserve Governor Lael Brainard recognized the dynamic of FIs working with third-party vendors and stated in a speech:

> Importantly, the guidance recognizes that not all aspects of a model may be fully transparent, as with proprietary vendor models, for instance. Banks can use such models, but the guidance highlights the importance of using other tools to cabin or otherwise mitigate the risk of an unexplained or opaque model. Risks may be offset by mitigating external controls like "circuit-breakers" or other mechanisms.

For example, because there is the potential for significant legal harms related to model risks associated with consumer credit underwriting use cases (e.g., bias resulting in a loan being declined), it would be expected that more extensive documentation be necessary to evidence the model design, theory, and logic. Alternatively, because the model output for general Risk AI/ML models is less likely to violate consumer protection laws and regulatory requirements, it is expected that the documentation associated with these models would be significantly less extensive.

Additionally, the MRM Guidance obligation to provide "public research" to support a particular use case could be interpreted to require a bibliography of academic findings or some type of documented recognition of an accepted industry practice for these models. However, such a requirement may be difficult or impossible to satisfy if the model makes advancements into new technologies or use cases.

For this reason, and to avoid discouraging innovation, extensive internal testing, a thorough effective challenge process, or dual validation processes for a particular time period should be confirmed as reasonable and acceptable ways to supplement what may be otherwise limited "public research" model validation evidence for new models. More specifically, if the rationale for requiring "public research" is to provide an independent source for model use validation purposes, this goal can be met by an effective challenge process requiring internal compliance personnel who are performing the validation to be independent of the model development and use process, and have no business stake in whether the model is deemed to be valid. Additionally, a dual validation process could require one of the validation processes to be performed independently.

Importantly, as noted above, demands for detailed access to vendor model development features and internal data can raise tensions between third-party model developers and banks due to intellectual property, security, and model integrity concerns, resulting in model innovation being disincentivized. Providing access to such information in the context of complex AI/ML models is also unlikely to advance core regulatory interests for the reasons noted above (including the greater importance of testing and outcomes for establishing safety and soundness of AI/ML models).

Instead, proper industry practice that can satisfy regulatory requirements and balance stakeholder interests includes the third-party vendor maintaining requisite documentation that provides the bank with appropriate transparency into the vendor development process related to the design/theory/logic of the model. This transparency would support the bank's ability to determine whether the third party maintains acceptable control and governance processes around an AI/ML model and would confirm the sufficiency of a model's design/theory/logic documentation.

Takeaway

Regulators could provide further clarity regarding documentation expectations with Risk AI/ML models, especially when the models are novel and/or developed by third-party vendors. The touchstone for the sufficiency of documentation should be what is needed for the bank to use and validate the model, as well as understand its design, theory, and logic. Detailed disclosure of proprietary information, including code, is unnecessary and unhelpful in verifying the sufficiency of a model and would deter model builders from sharing best-in-class technology with financial institutions.

D. Regulators should promote the development and recognition of industry practices and global standards

To promote more consistent compliance with the MRM Guidance, and broader understanding and application of its risk mitigation principles across all financial authorities, regulators should encourage industry participants (e.g., model developers and financial institutions) to develop standards and identify best practices related to the model development lifecycle. Regulators can foster this activity by recognizing published industry standards and best practices as compliant with regulatory expectations, as well as engaging with the industry in robust training and related forums.

Examples of non-regulatory standards development efforts related specifically to AI risk management both on the national and international scale include the NIST draft AI Risk Management Framework ("AI RMF") and the draft interim companion Playbook, and the International Organization for Standardization ("ISO") risk management standards.[29]

The AI RMF is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. The NIST Playbook includes suggested actions, references, and documentation guidance for stakeholders to achieve the outcomes for two of the four proposed functions in the AI RMF, with draft material for the remaining two functions to be released at a later date. The ISO 31000 is the international standard for risk management originally issued in 2009 by the ISO and revised in 2018. It provides a detailed framework for the design, implementation, and maintenance of risk management on a firm-wide basis and provides a guide for businesses to compare their existing practices with international standards.

Regulators could consider providing that adherence to guidance issued by NIST and ISO related to AI model risk management is presumptive evidence of compliance with the MRM Guidance. NIST, in particular, sets out pathways to enable the creation of specific Risk Profiles that allow for the industry to adapt risk management practices for specific contexts and use cases. By operating off foundational guidance to create the specificity that may be needed, we can ensure that best practices for risk management grow alongside new applications.

[29] NIST is a physical sciences laboratory and non-regulatory agency of the United States Department of Commerce. Its mission is to promote American innovation and industrial competitiveness. ISO is an independent, non-governmental international organization with a membership of 167 national standards bodies. See also, supra note 10.

Takeaway

Regulators should support the development of global standards that can be used across the financial services and regulatory landscape by explicitly recognizing such standards as presumptive evidence of compliance with the MRM Guidance and sound AI/ML risk mitigation practices, and by fostering further industry collaboration and training based on such standards.

E. Regulators should facilitate use of other governance controls that address the potential impact of AI/ML models, including incremental rollouts and circuit breakers

Equally important to testing protocols and validation checks are safeguards to limit the potential for negative impacts of models once implemented. Two well-accepted safeguards in the context of AI/ML models include incremental rollouts of models and the use of circuit breakers or check points. The purpose of incremental rollouts is to contain the model output and possible negative impact to the business goal of the model by validating model output incrementally rather than after the expected longer-term model purpose or use is fully implemented. This allows interim model modifications to be made if inaccurate model output is produced.

Additionally, circuit breakers permit models to be immediately suspended when metrics are triggered identifying model output that suggests the model is not operating as intended. Given the potential operational impact of a "kill switch," banks must have processes in place regarding such action to minimize the temporary sidelining of a model, including through the development of fallback operational plans.

For both incremental rollout and circuit breaker strategies, banking authorities can encourage further adoption by confirming that such practices used in the development and implementation of Risk AI/ML models are an "MRM best practice" for risk mitigation. These tools could be prioritized with AI/ML models and regulators could consider further industry collaboration and guidance in developing and sharing related best practices.

Takeaway

Regulators should use guidance to advance the use of governance controls, including incremental rollouts and circuit breakers, as key available tools in mitigating risks associated with Risk AI/ML models.

6. Additional considerations

A. Examiner training

Use of the MRM Guidance will only produce effective, real-life results if bank examiners, and financial regulators more broadly, are trained to implement the MRM Guidance as consistently as practicable. Accordingly, training of examiners and regulators should be conducted to ensure the development of consistent compliance expectations and to reduce the risk of idiosyncratic applications of the Guidance by individual examiners. Training should be conducted across conduct, financial markets, and banking regulatory agencies, including potentially as part of the uniform procedures and training of the Federal Financial Institutions Examination Council (FFIEC), to ensure a common baseline. It should also include the review of standards development efforts, including those noted above.

B. Industry training & forums

The need for technology and regulatory literacy regarding AI/ML models extends beyond the financial regulators to the financial institutions they regulate and the third-party model developers with which such institutions engage. It is increasingly important for financial institutions to have the internal expertise to integrate and monitor AI technologies and for developers to embed compliance. Accordingly, it is important for financial institutions to invest in their in-house technical expertise and capabilities and ensure their company leadership and model users have sufficient technological literacy. Similarly, third-party model developers need to deepen their knowledge and expertise regarding regulatory expectations, including with respect to compliance with the MRM Guidance. To this end, we encourage regulators to pursue joint industry and regulator training and collaborative forums to further raise the baseline of shared knowledge.

C. Fine-tuning the MRM guidance

In April 2021, the Federal Reserve, the FDIC, and the OCC issued the Interagency Statement on Model Risk Management for Bank Systems supporting the Bank Secrecy Act/Anti-Money Laundering Compliance to provide more tailored guidance related to models used for anti-money laundering purposes. This guidance has set a precedent for banking authorities to iteratively refine their regulatory approaches as necessary. An opportunity exists for the banking authorities to issue similar refined guidance premised upon industry best practices that could take the form of annotations to the MRM Guidance or through developing safe harbor frameworks that encourage certain behaviors. As noted in the sections above, such additional guidance could assist and guide model developers and users, and advance more consistent model practices.

7. Conclusion

Advances in AI/ML technology hold substantial promise for the future of banking and financial services. These very same technologies could also transform how financial regulators supervise activities and markets, and safeguard consumers. While we commend regulators for providing a sound framework in existing MRM Guidance for identifying and mitigating potential risks posed by AI/ML models, more can be done to increase certainty, clarity, and effective and efficient risk mitigation strategies.

This paper has focused on exploring the application of existing MRM Guidance to Risk AI/ML models used by financial institutions to identify potential fraudulent, illicit, and otherwise problematic financial activity. Our goal has been to suggest areas for incremental improvement of the shared understanding between regulators and the industry on expectations and best practices for mitigating risks associated with Risk AI/ML models.

As a threshold matter, we suggest that regulators further clarify and underscore that the mere use of AI/ML technologies does not inherently make a model high risk. We also urge regulators to recognize the importance of robust testing on Risk AI/ML models relative to a focus on explainability for purposes of establishing safety and soundness of Risk AI/ML models.

With respect to MRM Guidance regarding proper documentation requirements, we urge regulators to identify approaches that take into account important intellectual property considerations and relative responsibility allocations, especially in the context of third-party vendor relationships. While the OCC has provided further parameters regarding documentation requirements, more can be done to reflect the state of the technology and realities of today's marketplace.

In addition, we stress the importance of fostering the development and adoption of global standards, as well as reliance on well-understood governance controls that further mitigate risk. On the former, we suggest that regulators explicitly recognize published industry standards and best practices that can demonstrate compliance with regulatory expectations. These efforts include, for example, NIST and ISO standards being developed in the context of AI/ML. On the latter, regulators should explicitly recognize certain governance controls, including incremental model rollouts and circuit breakers, as capable of significantly mitigating risks associated with AI/ML models.

Lastly, we further urge ongoing collaboration between regulators and financial institutions whether in the context of training or sharing information. As noted above, the fast pace of development of AI/ML technologies requires constant education—among industry participants and regulators alike—and information-sharing regarding best practices and marketplace developments. We also encourage regulators to recognize that engagement and guidance in this space cannot remain static and instead must reflect the dynamism of the technology and the opportunities it presents.