Managing the risks of machine learning and artificial intelligence models in the financial services industry

pwc

Contents

1 Introduction

Rapid proliferation of machine learning algorithms across various business areas is exposing financial institutions to certain incremental risks that require bespoke governance and control mechanisms to manage. The purpose of this article is to provide our views on how the financial services institutions can build upon their existing Model Risk Management frameworks to effectively and efficiently manage these risks.

2 Background

Since the early 2000's, following the introduction of the Office of Comptroller of Currency's first guidance on Model Risk Management – OCC 2000-16 – banks and other financial institutions in the US have been steadily building governance frameworks, infrastructure, and teams necessary to manage risks associated with their use of models. A more comprehensive and expansive guidance was issued in 2011 jointly by the Office of Comptroller of Currency (OCC 2011-12 bulletin) and the Federal reserve (SR 11-7 bulletin) which accelerated the buildup and maturation of the Model Risk Management ("MRM") frameworks and their expansion to a much broader range of models used across the entire enterprise.

OCC 2011-12 / SR 11-7 requires that banks inventory models used across every business and functional area regardless of how simple or complex, and establish mechanisms for managing model risks across the entire model lifecycle. "Model risk" is defined as the possibility of financial or reputational loss resulting from the use of conceptually unsound or otherwise defective models, as well as inappropriate use of sound models.

A "model" under the regulatory guidance is defined as a computational system that produces numerical outputs or business decisions where the outcome is inherently uncertain. Such uncertainty can be the result of using complex statistical and other types of mathematical/analytical methods or it can be due to the reliance on judgmental assumptions. In other words, a system that simply applies a set of arithmetic calculations or deterministic rules is generally not classified as a "model".

It is important to note that none of the definitions of a "model" used by the regulators or by industry participants are capable of making an absolutely certain "black or white" determination of whether a computational system is a model or not in every single case. There are always some systems that MRM practitioners consider to fall into a "gray area" where the institution's MRM team has to make a judgment call. As the machine learning ("ML") and artificial intelligence ("AI") techniques find their way into an ever broader set of functional areas, we expect some of the ML/AI systems to fall into such a gray area[1]. However, by far the majority of ML/AI systems are very clearly identifiable as models under existing model definitions due to their mathematical/algorithmic properties (that result in the presence of the above-mentioned uncertainty) and direct business use. This involves, for example, systems used for marketing purposes, credit underwriting, and fraud detection that rely on techniques such as neural networks, gradient boosting, and random forest.

Please note that, even though a typical ML textbook includes traditional statistical regression and basic decision trees in the category of machine learning algorithms, for the purpose of this paper, we define ML and AI techniques to only include those that are considered to be mathematically and algorithmically complex and more "advanced" than the traditional commonly used techniques. This includes, for example, neural networks, natural language processing, and ensemble methods, such as, for example, random forests or gradient boosting.

[1] For example, some of the applications of Natural Language Processing techniques as well as customer care chatbots (e.g., those that are based on simple slot filling or regex) have caused discussions of whether these systems should fall into the scope of Model Risk Management team reviews.

3 Machine learning/AI risks

While many of the risks associated with the use of ML/AI-based models fall into the traditional model risk categories that a typical MRM framework is already designed to handle, there are some incremental risks that should be considered. In addition, some of the traditional model risks are greatly amplified for ML/AI-based models. Specifically:

Transparency

The vast majority of the existing and emerging ML/AI techniques (as defined in the previous section) offer reduced transparency into the inner workings of the algorithms. For example, while a typical statistical regression model can be represented by a simple equation that clearly shows what data inputs impact the final output, in which direction, and to what extent, no such clear representation is possible for many of the more advanced ML/AI techniques. Note that the degree of opaqueness does vary significantly from one technique to another.

The lack of transparency impairs the ability of the model developers, independent model validators, regulators, and other parties to ensure that the relationships captured by the model algorithm are conceptually sound and do not simply represent spurious correlations found in the input data that will not persist in the future.

Explainability

For some of the business applications of models, it is critically important to explain what specific inputs caused a particular outcome and how the outcome can be improved by modifying the inputs in question. For example, when a customer is rejected for a credit card, the lender must be able to explain which of the customer characteristics significantly contributed to this decision. These "reason codes" would inform the customer that they may be able to get approval in the future if, for example, they applied for cards less frequently than before, or paid off some of their outstanding balances, etc.

It is not just the lack of transparency that decreases the degree of explainability of some of the ML/AI model outputs, though it is certainly a major contributor. Outputs from many of the **traditional** algorithms vary monotonically as you vary inputs one at a time. For example, your credit score based on a well-designed Logistic regression would typically always go up as your length of credit history goes up. The advantage of many ML/AI algorithms (from the predictive accuracy perspective), on the other hand, is that they excel at identifying nonlinearities and non-monotonicities in the training data and, if left unchecked, would not naturally allow for monotonic-only relationships in the final specification[2].

Bias and Fairness

The old adage of "garbage in – garbage out", while not unique to ML/AI-based systems, is frequently greatly amplified by the use of ML/AI algorithms. One reason relates to the above-mentioned ability of the ML/AI algorithms to better identify and capture nonlinear and non-monotonic relationships in the training data, which sometimes leads to more frequent capture of spurious correlations compared to traditional techniques. The other reason is that ML/AI algorithms are most commonly used in conjunction with very large volumes of data compared to the datasets typically used to develop more traditional statistical models. The size of the training datasets, and especially in the context of leveraging and combining data from multiple disparate sources (including "alternative" data sources), is typically directly correlated with the likelihood of undetected data errors.

[2] An interesting side note is that some industry participants and vendors are experimenting with a sub-class of ML/AI algorithms that are restricted to produce monotonic relationships to overcome this challenge.

These and other reasons increase the risk that predictions or decisions from the ML/AI-based systems would be biased in some fashion. One specific type of such bias risk relates to fair lending regulations. While it is fairly easy for the developers to exclude from the training data the types of inputs that are known with certainty to be prohibited, e.g., gender, race, ethnicity, and age, ML/AI-based systems are more likely to be able to identify other inputs that can serve as proxies for the protected categories.

When thinking about the bias risks, it is helpful to explicitly consider how we are defining bias and fairness. Bias occurs, for example, when we discriminate against (or promote) a defined group consciously or unconsciously, and it can creep into an ML system as a result of skewed data or an algorithm that does not account for skewed data. For example, an ML system that reviews job applicants by learning from a company's historical data could end up discriminating against a particular gender or race if that group were underrepresented in the company's hiring in the past.

Fairness, meanwhile, is a social construct. And in fact, when people judge an algorithm to be "biased," they are often conflating bias and fairness: They are using a specific definition of fairness to pass judgment on the algorithm. There are at least 20 mathematical definitions of fairness, and when we choose one, we violate some aspect of the others. In other words, it is impossible for every decision to be fair to all parties.

Overfitting

The risk of model overfitting -- that is, the risk that the model would work well on the training data, but will break down when new data is consumed -- is amplified by the use of ML/AI algorithms. In traditional statistical modeling, this risk is typically controlled by reducing the number of features included in the model, and carefully reviewing how each feature enters the model equation from the perspective of whether the relationship makes conceptual sense. Deep learning methods, on the other hand, take away 'feature engineering' from human judgement. This aspect coupled with the relatively greater opaqueness of ML/AI-based systems can impair the reviews of model specifications for conceptual soundness. Additionally, it is much more common for an ML-based system to simultaneously rely on hundreds of features, which naturally increases the risk of overfitting.

Democratization

It is becoming exceedingly easy for someone who is not a part of the institution's analytics department and is not formally trained in statistics or data sciences to download the widely available Python or other programming code libraries, or fully integrated ML/AI vendor modeling software solutions, and build fully-functional ML/AI models with minimal effort. Such individuals may not be aware that what they are creating constitutes a model that should be subject to the institution's MRM framework, and may bypass all the governance and control mechanisms to use such model to make business decisions.

Moreover, although data scientists may possess the required technical skills for model development, they may lack domain knowledge and understanding of the data generation process associated with the specific business problem of interest. For example, ordering images by the likelihood of depicting a cat is very different from ordering customers by the likelihood of defaulting on a mortgage loan. Lack of domain knowledge coupled with opaqueness of the model mentioned above further amplifies the risk that the model is based on spurious correlations and may produce an erroneous, unstable, or otherwise unusable result. Some industry practitioners believe that the next major blowout of a model at a bank will be a result of such a scenario.

Change Management

The majority of the ML/AI-based models used in the financial services industry today are "static", meaning that, once the model specification has been developed and put in production, the model rules and parameters remain static until explicitly modified by the developer, which may not occur for months or years. However, we expect to see progressively more "dynamic" models going forward where embedded algorithms self-recalibrate frequently without human interference, sometimes daily or hourly.

This includes the class of continuous learning models where this improvement process never stops.

Use of more dynamic algorithms of this type presents increased risks of inappropriate changes that may result in model performance deterioration or in the introduction of the above-mentioned biases. A typical MRM framework incorporates rigorous change management controls for static models, and is likely ill-equipped to handle the risks of unsupervised (or minimally supervised) frequent model recalibrations.

Data Quality Risks

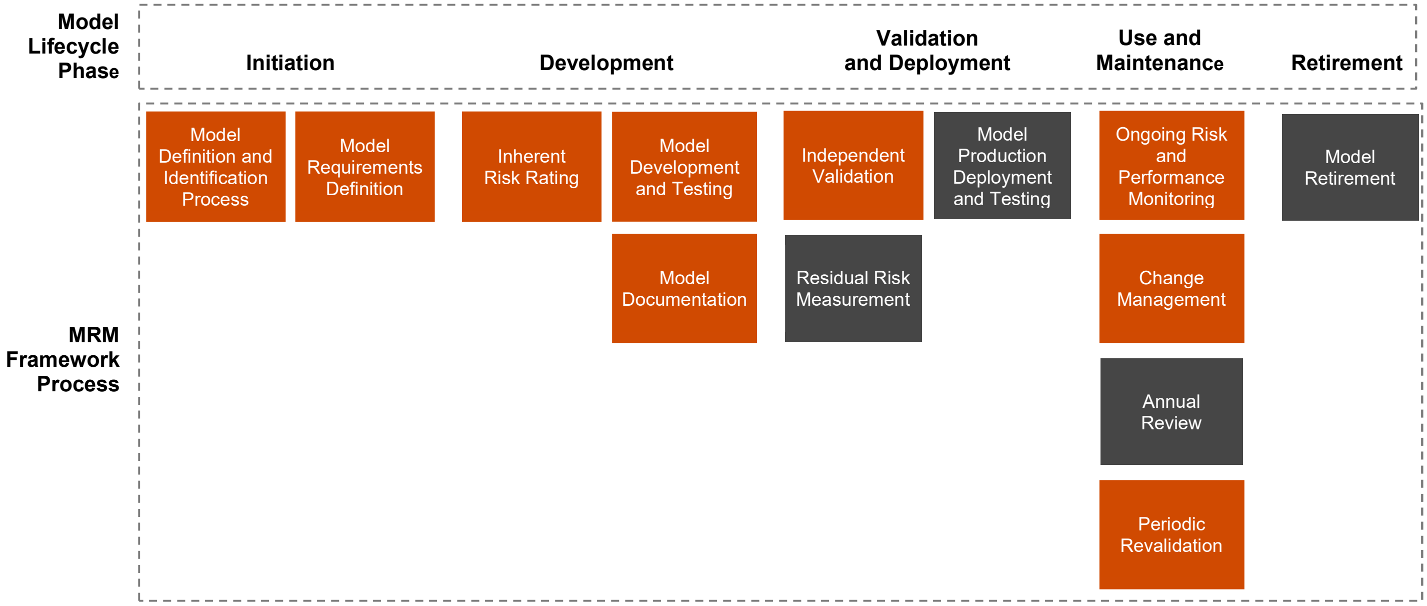
One of the advantages of ML/AI-based models over traditional models is the ability to process large volumes of observations and potential explanatory factors/model features as part of the model training. With this ability comes an increased burden of enhanced data governance. Financial institutions are expected to closely monitor and assess the quality of any data uses as part of the model development, especially for models used in credit underwriting, stress testing, and regulatory reporting. Over the past decades, institutions have heavily invested in data lineage projects to ensure that sources and uses of data are well understood and documented. However, historically, in ML/AI model development, data quality has taken a back seat, with key considerations being privacy and confidentiality.

Open Source Software Risks

A decade ago, the majority of statistical and mathematical model development has been undertaken using “industry standard” commercial software packages such as SAS, STATA, MatLab, and others. With large dedicated teams of software engineers at each of the analytical software publishers, a large customer base, fairly stable code base with infrequent new version releases, and centralized and formalized defect fixing process, the risk of inherent defects within the model development software was considered to be minimal. The recent explosion of the number of available open-source analytical packages, including R and Python, brings with it an increased risk that the tools used to develop models themselves may include defects or sub-par algorithms that may lead to development of sub-optimal or downright erroneous models.

4 Recommended model risk management framework enhancements

A typical industry MRM framework consists of the following key controls and processes covering the full model lifecycle:



Model Lifecycle Phase	Initiation	Development	Validation and Deployment	Use and Maintenance	Retirement
MRM Framework Process	🟧 Model Definition and Identification Process 🟧 Model Requirements Definition	🟧 Inherent Risk Rating 🟧 Model Development and Testing 🟧 Model Documentation	🟧 Independent Validation ⬛ Model Production Deployment and Testing ⬛ Residual Risk Measurement	🟧 Ongoing Risk and Performance Monitoring 🟧 Change Management ⬛ Annual Review 🟧 Periodic Revalidation	⬛ Model Retirement

The colors in the above chart signify whether the particular process or control is impacted, in our view, by the introduction of ML/AI models (orange) or not (gray).

A typical MRM framework in the industry consists of a number of documents and artifacts:

- A Model Risk Management policy, which tends to be short and lays out the MRM principles, key controls, and responsible stakeholders and committees;
- Detailed MRM standards and procedures; and
- Supporting templates and tools.

In our view, most of the proposed enhancements to the MRM framework should be implemented in the MRM standards/procedures and templates/tool as opposed to the MRM policy. The following sections dive into each of these processes and detail our recommendations for the MRM framework enhancements.

Please note that we expect some of the risks discussed earlier, specifically those of bias and fairness, to also be addressed outside of the MRM framework through a dedicated AI/ML governance program.

Model Definition and Identification Process

While the model identification process is not really different for ML/AI systems, the above-mentioned risk associated with the democratization of ML/AI tools needs to be considered. Institutions typically have requirements for any staff involved with models to take training on the institution’s MRM framework to ensure that they understand the risks that models can pose as well as their responsibilities for managing these risks. However, the challenge now more than ever is to ensure that such training reaches a broader audience of staff that may potentially decide one day to download and use ML/AI packages and tools. In practice, this may mean exposing nearly the entire staff of the institution to some minimal MRM training.

Besides the preventative controls like training, some institutions are also considering deploying automated detective controls, including, for example, scanning employees' computers for presence of common ML/AI packages and tools.

When it comes to model definition, the same SR 11-7 principles still apply, but we have observed a tendency in the industry to (1) conflate the concepts of "automation" and "machine learning / AI", and (2) extensively debate whether certain types of ML/AI applications (e.g., the chatbots) should really be classified as "models" and fall within the MRM team's jurisdiction. As such, we recommend adding specific language designed to address these and any other known areas of ambiguity related to the classification of ML/AI systems into the section of the MRM standards or procedures related to model definition and identification. Over the past decade, institutions have added similar types of qualifying language to deal with other "gray area" situations, such as expert-based/qualitative estimates, to help minimize ongoing debate and confusion.

Model Requirements Definition

A formal process for defining and documenting model business and technical requirements prior to the start of model development process (or prior to the evaluation of vendor solutions) is a requirement at some of the financial institutions. At other institutions, this part of the model lifecycle is subject to less formality, and the requirements are documented later as part of the model documentation. Due to the above-mentioned risks of ML/AI-based systems, institutions may want to formalize a gating process early in the model lifecycle that would, among other things, prevent usage of ML/AI techniques in business areas where the institution may consider such use to be highly undesirable. For example, due to the opaqueness of some techniques, their use in regulatory reporting areas (e.g., stress testing) or in financial reporting may be deemed inappropriate.

To avoid having developers go down the wrong path, an institution may want to establish an explicit ML/AI risk appetite framework coupled with an early gating process to ensure the developers are aware of the "dos and don'ts" of ML/AI modeling before embarking on a long and expensive development process. Such a framework may define, for example:

- Certain business and functional application areas where the use of opaque or unexplainable ML/AI techniques is generally prohibited,
- Other business and functional application areas where the use of these techniques is permitted, but with certain controls, and
- Business and functional application areas where unrestricted use of ML/AI techniques is permitted. For certain low-risk areas, this risk appetite framework may even permit less rigorous testing and documentation requirements.

We see such a framework being owned by the Chief Risk Officer to help ensure broadest applicability across the entire organization. The framework should align with the overall risk tolerance of the organization, but this should not lead risk taking institutions to neglect internal controls or regulatory considerations, such as fair lending or data privacy rules. Done properly, the ML/AI risk appetite framework will establish a structured approach to development, integrating control and regulatory considerations, with leadership at all levels carefully and periodically ensuring business units are appropriately applying it during model planning and development.

Inherent Risk Rating

A typical industry model inherent rating/tiering framework determines the rating based on three key considerations:

1. The model's business use (e.g., underwriting, financial reporting, regulatory reporting, marketing, etc.),
2. The model's "materiality", and
3. The model's complexity.

While the first two dimensions are not impacted by whether a model is based on an AI algorithm or a more traditional modeling methodology, measurement of "complexity" certainly is. ML/AI algorithms tend to be more complex to develop and implement and, as discussed earlier, typically offer reduced transparency into the model mechanics, which includes the likelihood of an undetected model issue.

As such, we propose to capture these increased risks explicitly in the section of the procedures/ standards (and associated risk rating tools) describing the methodology for measuring the model's complexity.

An institution may also want to consider adding another dimension to the risk rating scheme to capture the degree of explainability of the outputs, where such explainability is deemed important in the context of the model's business use.

Model Development and Testing

By far, this aspect of the MRM framework is impacted the most, in our view. Strong and mature MRM frameworks in the financial services industry incorporate requirements for the formalization of detailed technical procedures and standards for developing and testing different types of models. We frequently see such detailed procedures and standards for stress testing time series models, credit models, financial instruments pricing models, AML/BSA and fraud models, etc.

Regardless of whether the development procedures/standards are owned by the first or second line of defense, such documents help ensure that models of a particular type are developed to the same quality standards and using consistent methods across different teams and individuals within the institution. By the same token, and especially given the increased risks associated with the model transparency, explainability, and bias, putting in place detailed technical procedures and standards for developing ML/AI models can go a very long way to ensuring these risks are appropriately mitigated in the early phases of the model lifecycle.

In our view, the technical procedures and standards may need to be differentiated by the model's intended business use (e.g., for marketing vs. underwriting or risk monitoring) and should incorporate the following key elements:

- Discussions of pros and cons of different methodologies. Per earlier discussion in the Model Requirements Definition section, which also covered the ML/AI risk appetite and integration with controls and regulations, an institution may deem certain methodologies to be prohibited for certain types of business uses. Outside of such basic rules, certain techniques are better suited for some applications compared to others.
- Discussion of approaches and minimum requirements for analyzing input data quality, including methods to detect inherent data biases. Use of non-traditional data sources and data labeling risks should be explicitly considered.
- Discussion of the preferred methodology for feature selection and candidate models' development.
- Detailed description of the types of testing that should be performed (specific to certain methodologies as well as common across all methodologies) to ensure that the model is conceptually sound, robust and stable, technically/statistically sound, and offers strong performance. For each test, the standard should describe:
  - The purpose of the test,
  - Details of how the test should be implemented,
  - **A priori** expectations for the test outcome, as well as a discussion of the risks associated with the test's failure and what actions should be taken to mitigate such risks (e.g., additional/alternative tests),
  - Description of how the test results should be presented and discussed within the model documentation, together with illustrative examples.

  This would include testing designed to directly mitigate the transparency and explainability limitations of the selected methodology, such as building partial dependence plots, proxy testing, or adversarial testing, for example.
- Discussion of requirements for obtaining reviews and approvals by other teams/departments, such as fair lending compliance reviews.

- Where open-source software packages are used, requirements for appropriate due diligence to ensure that these packages are free from significant defects. For example, parallel testing using alternative software packages that have been previously deemed reliable is commonly used for such due diligence. More broadly, the organization's MRM and IT departments may wish to establish a formal program for vetting analytical software and model development tools coupled with controls preventing employees from downloading and installing unapproved tools.

Model Documentation

A leading industry practice is to maintain model development standards and templates that cover common elements across all models types, but also include content that is specific to particular model types. For example, it is quite common to have separate documentation template and guidance for statistical models vs. financial engineering models vs. AML/BSA/fraud models. Such decisions for differentiated templates/guidelines are a direct outcome of the recognition that the processes for developing and testing different model types can be significantly different. For example, detailed description of the variable selection process is typically a key component of documentation for statistically-estimated credit loss models, but is not relevant for options pricing models developed using the Black Scholes formula.

Using the same logic, we consider it to be beneficial to evaluate whether a separate documentation template and guidance for ML/AI-based models may be warranted, or whether one of the existing templates (e.g., for statistically estimated models) could be enhanced to capture the more unique aspects of such models.

Independent Validation

Just like the need for the specialized technical model development and testing procedures and standards for the 1st line, we feel a similar ML/AI-specific technical validation standard is needed for the 2nd line. Mature and effective model validation functions in the industry usually maintain a library of such testing standards, procedures, and detailed testing plans for a broad range of models they commonly validate. Adding an ML/AI-specific set of validation standards/plans is a natural extension of their validation framework and is key to ensuring consistent and high-quality validation of every ML/AI-based model across different sub-teams, individuals, and over time.

Change Management

As mentioned earlier, we see unique challenges for managing changes for those ML/AI-based models that are considered to have "dynamic" structure and/or parameters and are subject to ongoing unsupervised recalibration. At many financial institutions, the model change management process already includes a "pre-authorized change" mechanism that works as follows:

1. For models that require frequent parameter recalibrations, the model owner may request that such recalibrations are excluded from the independent validation by the institution's Model Risk Management team.

2. During the initial model validation, the MRM team will review the request and either approve or reject it.

3. If approved by MRM, certain types of ongoing changes, such a parameter retuning, can be done by the model owner without submitting a model change request to MRM. Typically, strict guardrails are put in place on such updates, including for example:

   a. Recalibration must not alter the model structure, including the variables/features included in the model, or their transformations;

   b. The same process for recalibration/tuning that was originally reviewed by the MRM team must be followed without deviation for subsequent recalibrations;

   c. Results of the recalibration must be compared to the previously-validated model specifications to ensure that the updated parameters do not deviate too far from the original ones;

   d. Performance of the recalibrated model must be compared to the performance of the previously-validated version to ensure that performance metrics improve or at least do not deteriorate significantly compared to the performance of the original model;

e. Recalibration process and results are captured in the formal model change log available to MRM team upon request.

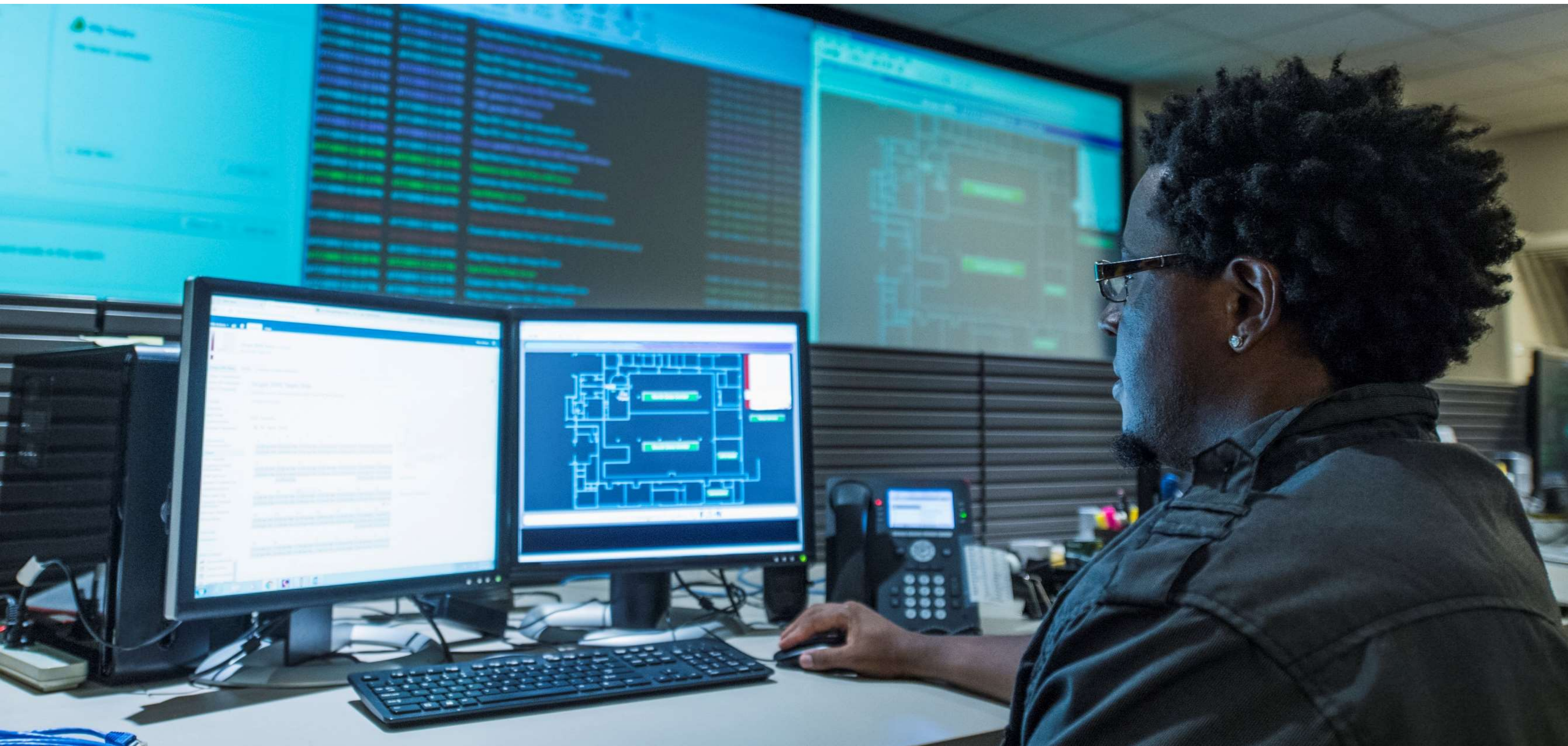
4. MRM team typically performs some assessment of recalibrations/tuning during the policy-required annual reviews or as part of ongoing model risk and performance monitoring process.

The same concept of pre-authorized changes can be applied to dynamic ML/AI-based systems. However, a number of the above-mentioned controls require automation to allow for unsupervised model updates. Certain triggers must be built into the automated recalibration and controls processes to escalate breaches in the parameter change and performance thresholds. We recommend that institutions implement enhancements to the change management sections of their existing MRM procedures and standards to explicitly cover the unique challenges associated with ML/AI-based systems, especially the dynamic ones.

Ideally, the control processes and notifications in the event of breaches would be built into the institution's centralized Model Risk Management inventory and workflow management technology platform to ensure robust capture, notifications, and escalation of exceptions to first line and second line of defense stakeholders. Such functionality does not currently exist in some of the MRM technology platforms used across the industry and will require thoughtful development.

Ongoing Risk and Performance Monitoring

Some of the key challenges for the ongoing monitoring of risks and performance for ML/AI-based systems relate to the change management challenges for dynamic systems detailed above. In addition, the greatest difference in monitoring ML/AI-based systems relative to traditional ones lies in the greater breadth of tests needed to evaluate these models stemming from the bias and overfitting risks discussed earlier. It is not enough to test for bias and overfitting just once during the model development; changes in the production data feeds coupled with model recalibrations can lead to rapid manifestations of these risks post-implementation. We recommend that detailed ongoing monitoring guidelines for different types of ML/AI-based systems are developed and incorporated into the above-mentioned technical development and testing standards.



5 Conclusion

Given the tremendous progress most banking and other financial services institutions made over the last 9-10 years towards effective model risk management, the incremental changes needed to properly address the unique or amplified risks resulting from the growing use of ML/AI-based systems are relatively less substantial. Nevertheless, the importance of such changes in preventing a major ML/AI model failure should not be underestimated. Our paper presents a clear roadmap for enhancing different aspects of existing Model Risk Management frameworks to help effectively manage these risks at every stage of the ML/AI model's lifecycle.

While this paper is targeted at financial services industry audience, we believe other industries have much to learn from financial services in designing the appropriate governance structures around models and the data that feed them. An effective governance is foundational to the deployment of Responsible AI across an organization, which translates the ethical landscape of an enterprise into concrete actions, and considers specific considerations for models like bias, explainability and interpretability, robustness, security, privacy as well as safety of systems. For more insights on Responsible AI across all industries, please visit www.pwc.com/us/rai

Contact information

Jason Dulnev
Principal, Model Risk Management and Risk Analytics
Jason.Dulnev@pwc.com

Anand Rao
Principal, Global Artificial Intelligence
Anand.S.Rao@pwc.com

pwc.com

This content is for general information purposes only, and should not be used as a substitute for consultation with professional advisors.