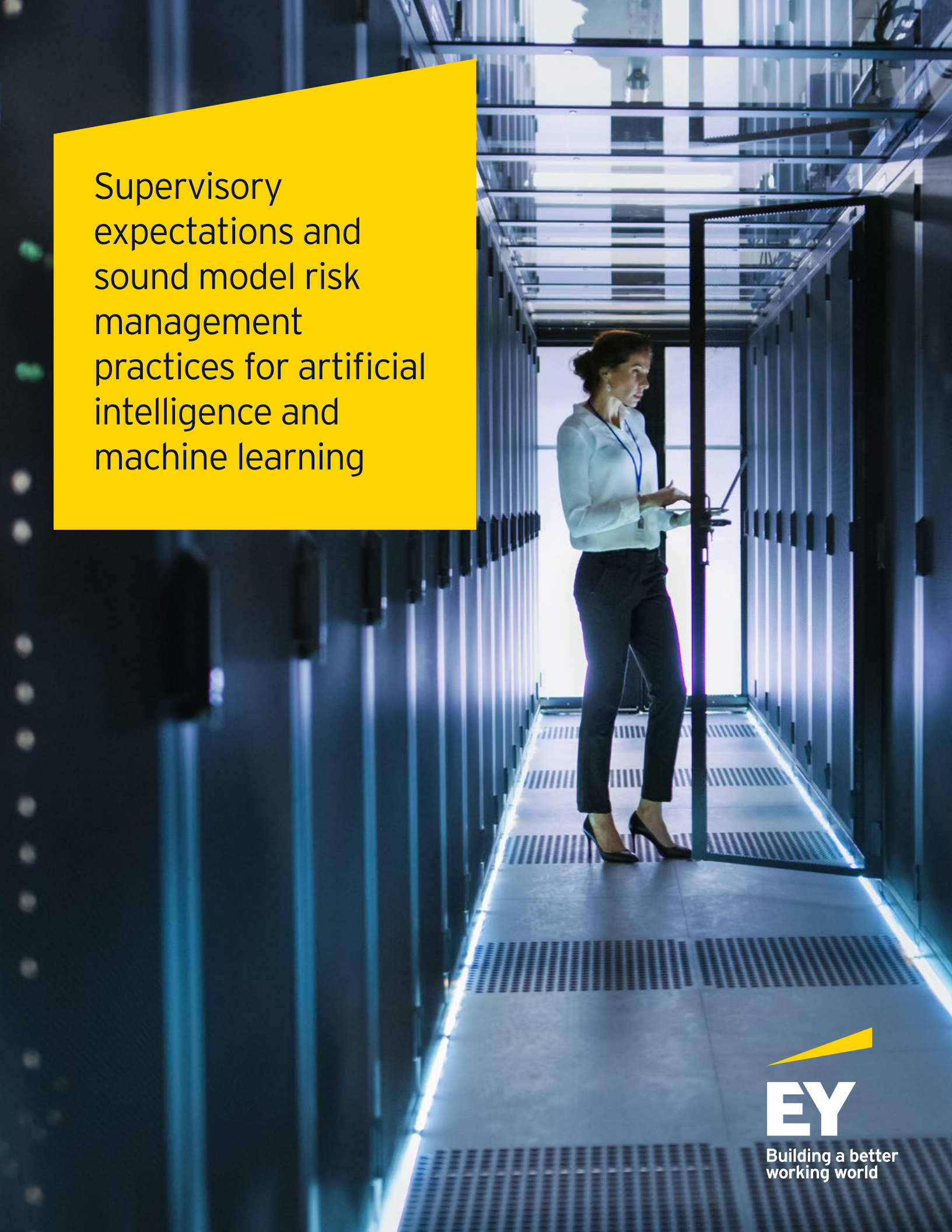
Supervisory expectations and sound model risk management practices for artificial intelligence and machine learning



EY
Building a better working world

1

Sound model risk management fosters accountability and enhances trust

Sound risk management of artificial intelligence (AI) and machine learning (ML) models enhances stakeholder trust by fostering responsible innovation. Responsible innovation requires an effective governance framework at inception and throughout the AI/ML model life cycle to achieve proper coverage of risks.[1]

Effective model risk management (MRM) is part of a broader four-step process to accelerate the adoption of AI/ML by creating stakeholder trust and accountability through proper governance and risk management.[2] These steps include:

- Developing an enterprise-wide AI/ML model definition to identify AI/ML risks
- Enhancing existing risk management and control frameworks to address AI/ML-specific risks
- Implementing an operating model for responsible AI/ML adoption
- Investing in capabilities that support AI/ML adoption and risk management

Effective MRM can further enhance trust in AI/ML by embedding supervisory expectations throughout the AI/ML life cycle to better anticipate risks and reduce harm to customers and other stakeholders.[3] This entails holding model owners and developers accountable for deploying models that are conceptually sound, thoroughly tested, well-controlled and appropriate for their intended use.

Financial services firms can in many respects leverage existing MRM processes, such as risk assessment, validation and ongoing monitoring, to address AI/ML-specific risks and align with supervisory expectations because the risks of AI/ML models are similar to those of more traditional modeling techniques. Nevertheless, four aspects of AI/ML will likely require additional investments in capabilities to align with current expectations. These include the growth in diverse use cases (e.g., document intelligence, advertising/marketing), reliance on high-dimensional data and feature engineering, model opacity, and dynamic training.

---

[1] We use the terms AI and ML interchangeably throughout for ease of exposition but in practice there are significant differences, with ML being a subset of AI. See https://www.fsb.org/wp-content/uploads/P011117.pdf.

[2] "Four steps to accelerate adoption of AI/ML in the US banking industry," *EY website*, https://www.ey.com/en_us/financial-services/four-steps-to-accelerate-adoption-of-ai-ml-in-the-us-banking-industry, accessed March 2020.

[3] "How can risk foresight lead to AI insight?" *EY website*, https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/advisory/ey-how-can-risk-foresight-lead-to-ai-insight.pdf, accessed March 2020.

These aspects of AI/ML will require greater investment in data governance and infrastructure and key elements of model life cycle risk management, including model definition, development and validation, change management and ongoing monitoring. These aspects will also require tighter linkage among the MRM framework, data governance and other risk management frameworks such as privacy, information security and third-party risk management.

Firms recognize the importance of trust in AI/ML and the need to enhance MRM capabilities to address the unique risks and aspects of AI/ML. According to the EY/IIF 2019 Global Risk Survey, firms plan to enhance their MRM processes for AI/ML over the next three years across a range of areas, including model risk assessment, ongoing monitoring, change management, and policies and procedures.[4] The survey also highlights that firms are making additional investments to enhance their broader AI/ML governance practices beyond MRM.[5] These investments highlight that firms recognize the need for broader governance mechanisms to effectively monitor and control other non-model-related risks arising from AI/ML such as privacy, information security and third-party risk.

US banking regulatory agencies are closely monitoring developments related to AI/ML. In their messaging and supervisory posture, US regulators are seeking to balance the benefits associated with innovation against the downside risks.[6] The balancing act they are striking is most evident in recent guidance regarding the use of alternative data in consumer credit.[7]

Based on public statements, we also understand that the agencies are developing revised guidance for AI/ML models, which may impact the scale and scope of investments in AI/ML capabilities. Nevertheless, the agencies have stated that existing MRM guidance remains relevant for the use of AI/ML models and believe that the principles underlying existing guidance provide a good basis for developing an effective MRM process for AI/ML. They have also stated that it is important for other risk management frameworks to consider non-model-related risks of AI/ML.

Our goal is to provide a road map for how to enhance MRM capabilities for AI/ML along the MRM life cycle that align with the principles underlying current supervisory expectations as reflected in existing guidance (e.g., SR 11-7/OCC 2011-12).

A key challenge for firms is how to apply the general principles of current expectations to the specific case of AI/ML in a manner that captures the inherent risks. More specifically, we explain the model risks associated with AI/ML models and describe enhanced capabilities that could address these risks with respect to development, use and implementation. We also explain the areas where firms can rely on existing processes for traditional models due to the similarity in risk.

---

[4] Global bank risk management survey 2019," *EY website*, https://sites.ey.com/sites/DS_BCM/Pages/global-bank-risk-management-survey-2019.aspx, accessed March 2020.

[5] Ibid.

[6] "Guidance for regulation of artificial intelligence applications," *White House website*, https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf, accessed March 2020.

[7] "Interagency statement on the use of alternative data in credit underwriting," *Federal Reserve website,* https://www.federalreserve.gov/newsevents/pressreleases/files/bcreg20191203b1.pdf, accessed March 2020.

Key challenges in applying SR 11-7 to AI/ML models

Distinguishing features of AI/ML models lead to challenges in applying SR 11-7

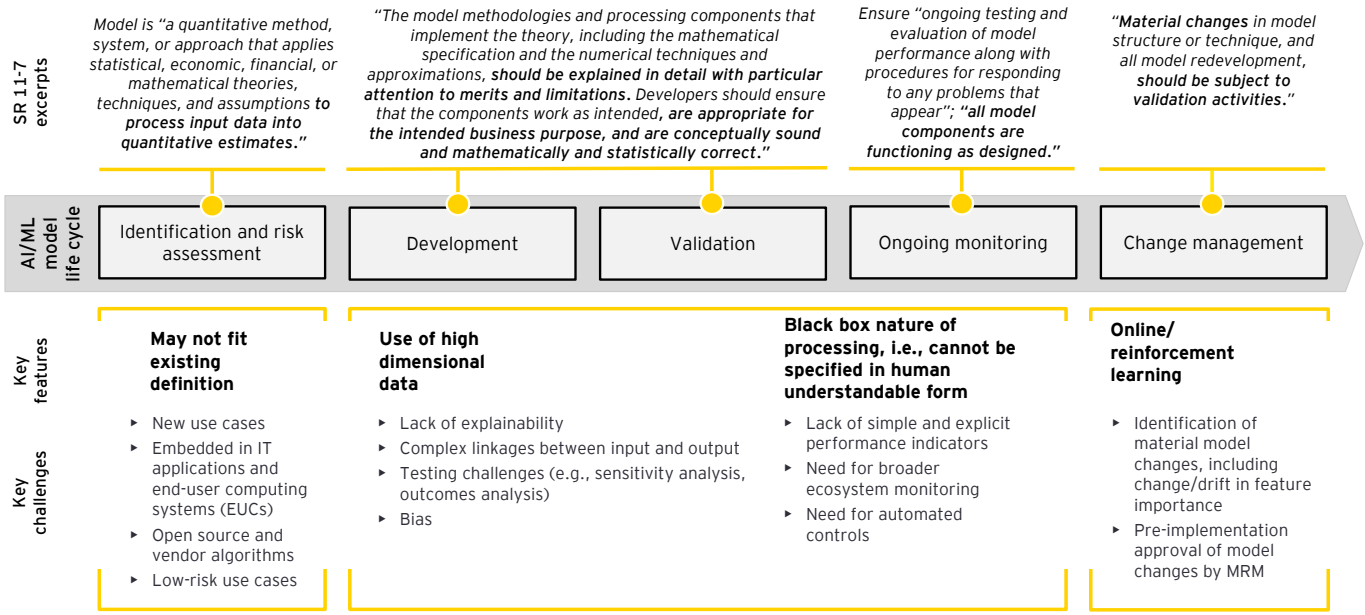
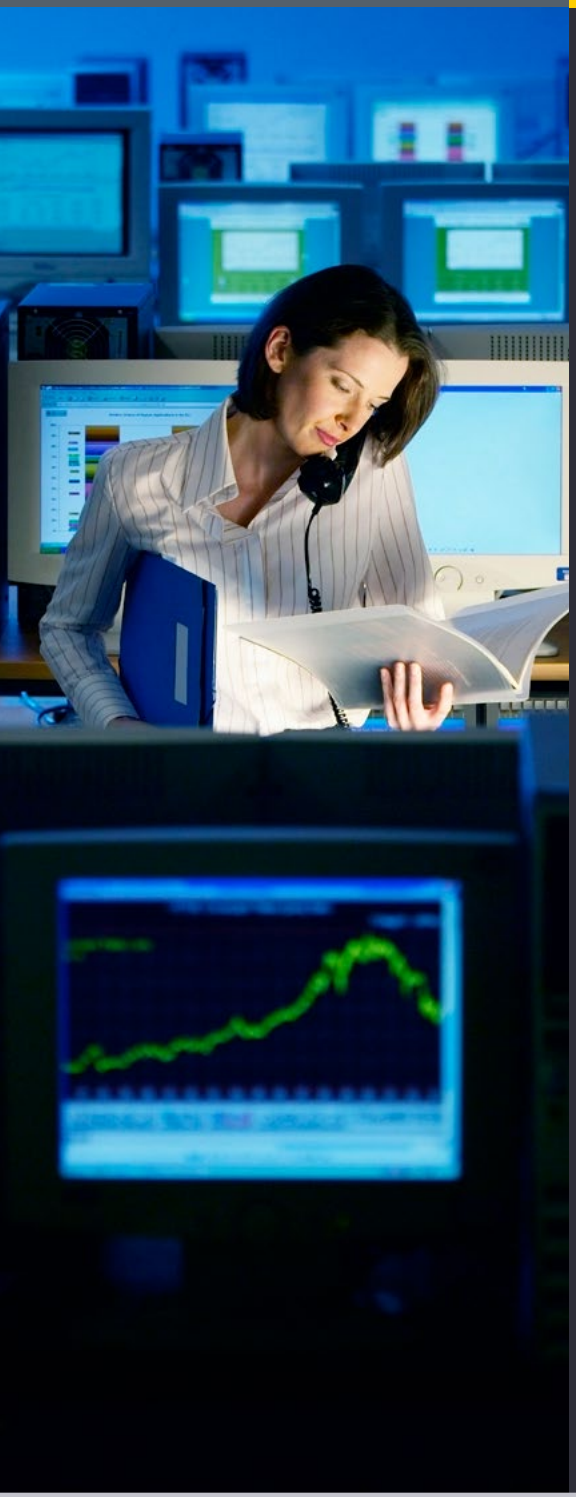


Figure 1: Applying existing supervisory guidance to AI/ML

2

AI/ML model risk is similar to traditional model risk, but risks can be more difficult to identify and assess



Like traditional models, such as logistic regression, AI/ML models, such as deep learning (DL), can expose a firm to risk because they can lead to adverse consequences and poor decisions if a model has errors in its design or construction, performs poorly or is used inappropriately. While the risks of AI/ML models are qualitatively similar to those of traditional models, the reliance on high-dimensional data, dynamic retraining, the opacity of the transformation logic and feature engineering can lead to unexpected results and make risks more difficult to identify and assess.

As with traditional models, poor performance can arise from implementation errors, including those related to calibration and poor data quality. In the case of AI/ML, the complexity of the model makes it more difficult to assess whether the results of the model can be generalized beyond the data used for training. The results may not be generally applicable if the model underfits or overfits the data in relation to a set of performance criteria.

Underfitting means that the model does not capture the data "well" in sample relative to the performance criteria. Overfitting means that the model fits the training data "too well" relative to a set of performance criteria and exhibits poor prediction performance when tested out of sample. As discussed in more detail below, poor data availability or quality can undermine model fit and lead to sampling bias and lack of fairness.

Also like traditional models, AI/ML models can be used inappropriately, giving rise to unintended consequences. The model result should be relevant and informative in understanding whether the desired business outcome is achieved. Risk can arise because the goal as defined by the algorithm is not clearly aligned to the real-world business problem statement. The intended use of the model also may not align with real-world applications due to issues noted later regarding data availability, quality and representativeness. As a result, the informativeness of the output to the business decision is overstated. Alternatively, the business goal that the algorithm quantifies may be aligned to the business problem, but it may not account for all relevant considerations, which can lead to unintended consequences, such as a lack of fairness.

Traditional statistical vs. AI/ML models
Key distinguishing features

Category	Traditional statistical models	AI/ML models
Model methodology	▸ Based on clear stochastic or statistical theory or assumptions ▸ Typically linear or transformed to nonlinear ▸ Use of limited number of explanatory variables/factors ▸ Use of quantitative method and business judgment	▸ Probability theory + structured model framework + engineering experience ▸ High dimensional and nonlinear ▸ Typically uses large number of data attributes (feature), which may not be always known ▸ Can use unstructured data
Data	▸ Use of low dimensional and structured data ▸ Measurable data quality standard	▸ Designed to handle large volume of high-dimensional data given its data mining nature ▸ Preparation of data and data labeling (for supervised learning) could be tedious, costly and time consuming
Model calibration/ training	▸ Standardized calibration procedures leveraging widely used optimization methodology (maximum likelihood estimation, LS, MM) ▸ Closed-form or semi-closed-form formulas may exist	▸ Model training is often dependent on the choice of hyper-parameters, layers, initialization, activation and cost functions and is a critical component in particular for a complex models (e.g., neural nets) ▸ Several methods to solve undertraining and overtraining (overfitting) issue, which turns training into an engineering problem
Implementation	▸ Use of in-house and vendor solutions ▸ Tractable replication is possible in most cases ▸ Lower demand on infrastructure – capacity, latency, compute	▸ Extensive use of open source and vendor algorithms and libraries, which may not be adequately documented ▸ Replication of model is at times difficult/not feasible ▸ Higher demand on infrastructure – capacity, latency, compute with increased use of cloud-based services
Model performance assessment	▸ Well-established statistical measures exist (p-value, R-squared, other parametric hypothesis tests) ▸ Explanatory variables/factors and attribution of results to them can be easily analyzed	▸ Model output is hard to assign to individual attributes ▸ Stability can be hard or infeasible to assess (especially for unstructured data) ▸ Theoretical performance tests could be hard or nonexistent (especially for unsupervised models)
Ongoing monitoring	▸ Well-established key performance indicators (KPIs) for modeling inputs, outputs and performance ▸ Models are retrained on an infrequent basis	▸ KPIs and thresholds are difficult to determine for models that utilize high-dimensional and unstructured data ▸ Models are frequently retrained to handle population shifts (online learning)

Figure 2: AI/ML models share similar attributes of traditional models on the surface but entail greater complexity

Unique features of AI/ML models

The problems to be learned are complex

- Typically the problems that are solved by ML models are complex and nonlinear.

Model specification is not formulated explicitly

- Machine learning models are generally fitting data in high-dimensional spaces (connected or disconnected), and representing such data in a manner that is understandable to humans is intractable or impossible.

Model training process

- Training data is typically high-dimensional, semi-structured or unstructured, and voluminous.
- Complex training method, such as online training, may be required.
- Extrapolation is hard to detect and avoid.
- Limited training data results in low-density regions in training space.
- Feature space is not uniform; more training data is needed for heterogeneous regions.
- Hard to define extrapolation vs. interpolation in the input data space.

Transformation logic: linkage between input and output

- Large number of variables used may create interpolation and/or extrapolation issues.
- Input may impact output through a nonlinear and/or non-monotonic relationship.
- Inputs may jointly impact the output, i.e., interaction terms exist.
- Inputs may be correlated (may lead to implicit bias, language terms can be correlated to gender, certain habits are correlated to geographic location).
- Use of nontraditional sources of data may increase.

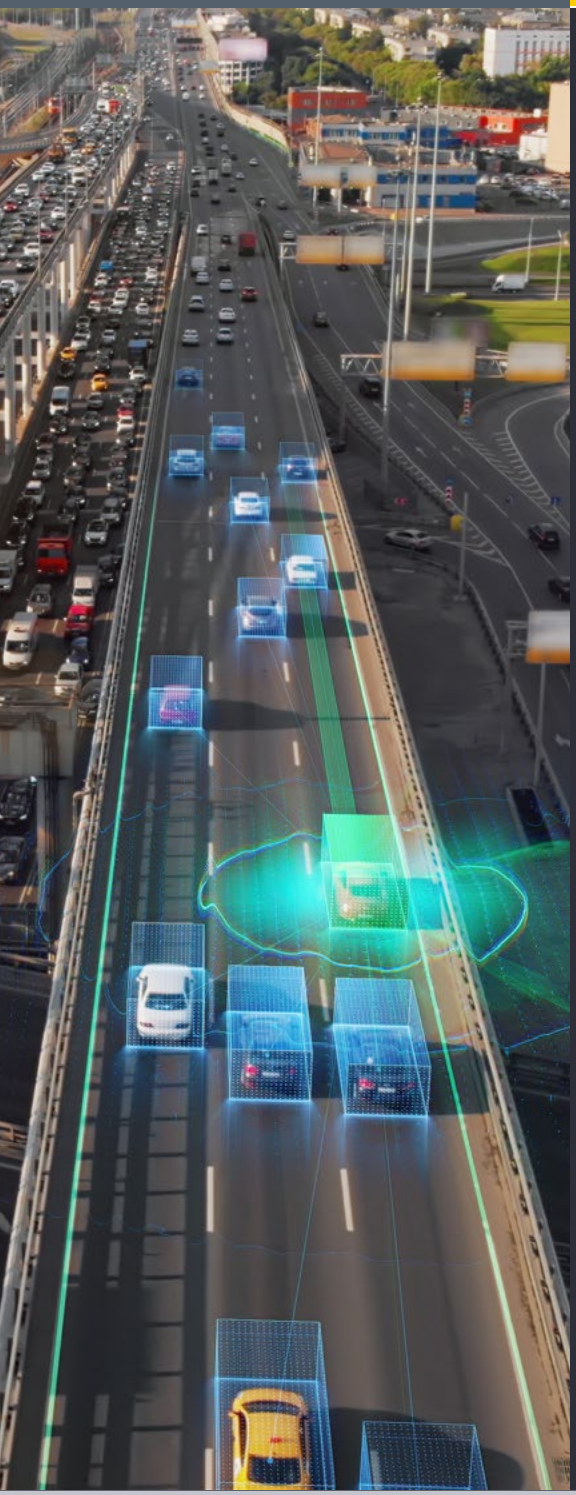
Emerging challenges

- Adversarial attacks by perturbating data not perceivable by humans while recognizable by machines.

Figure 3: The unique features of AI/ML models pose specific risks

3

Manage AI/ML model risk across the life cycle and enhance capabilities to reflect AI/ML attributes



A. Model definition standards

To identify AI/ML models, banks may consider leveraging the existing model inventory management process, augmented with specific considerations for certain AI/ML techniques, key characteristics of the algorithms (such as dynamic calibration) and other capabilities. Processes and capabilities would need to be enhanced by establishing standards that define whether various AI/ML models should be captured in the inventory as models requiring independent validation.

MRM should also provide training to model developers so they understand what needs to be reported to MRM. Standards are important because AI/ML models can be embedded within numerous software applications and processes, making them difficult to identify. Training for developers is important because AI/ML developers may be new to MRM expectations, especially since AI/ML models, such as those created by third parties, can be developed outside traditional development groups or channels.

Given the range of AI/ML techniques, platforms, vendors and capabilities, it is important to adopt consistent standards across the enterprise of what constitutes an AI/ML model. The standards should be embedded within innovation programs, new product/business approval processes, third-party sourcing, information technology (IT) software implementation and updates, and other relevant programs across the organization.

Even with standards adopted, developers may not be able to reliably or consistently identify AI/ML models because the range of use cases is so wide — for example, chatbots, natural language processing models for disclosure review, recommender systems used in marketing and more. As a result, MRM should periodically review the AI/ML development pipeline to confirm the inventory remains accurate and complete. MRM should also periodically ask developers to attest to their understanding of the criteria and reporting expectations to confirm that standards are consistently applied across use cases and over time.

MRM should also update the inventory by providing AI/ML-specific criteria based on AI/ML attributes. Updating the attributes is important so models can be traced to their underlying specifications and components as updates are made over time. The attributes can be included in the inventory as metadata and can include source code, data inputs and labels, features, explainability (if necessary), and retraining frequency.

MRM framework enhancements for AI/ML models

Model risk management framework must be enhanced to incorporate AI/ML-specific considerations

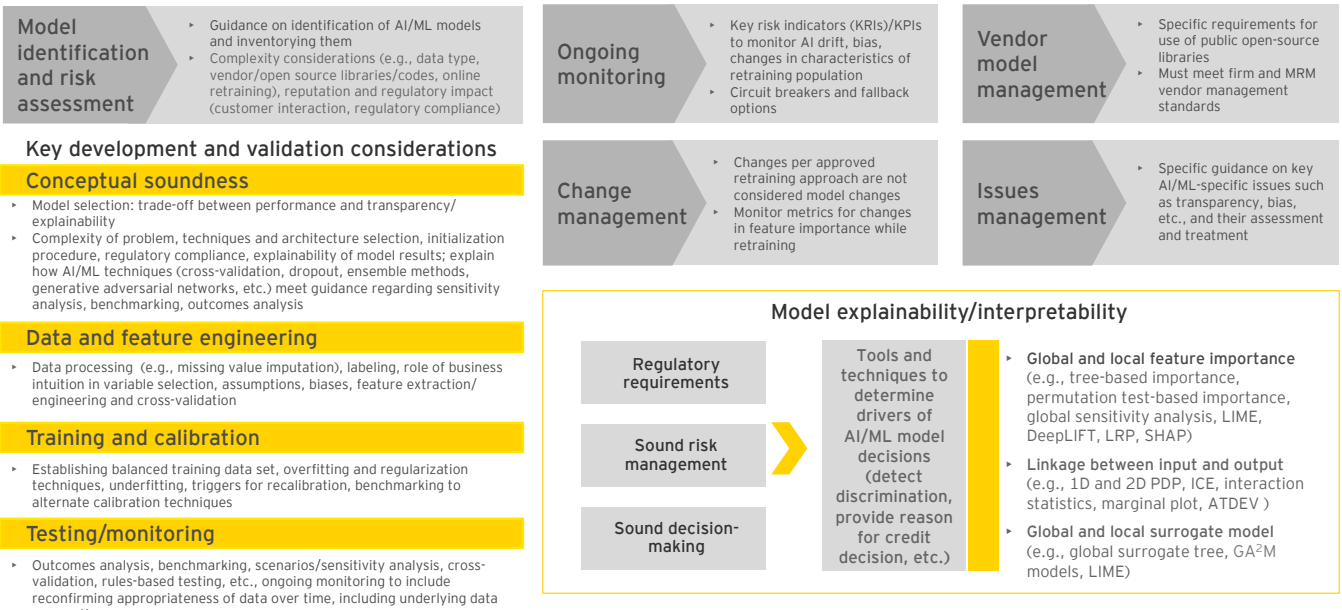


Figure 4: Enhanced capabilities are needed to capture the risks associated with AI/ML models

Are AI/ML applications models?

Regulators will likely rely on existing guidance in SR 11-7/OCC 2011-12 as necessary for assessing model risks of AI/ML applications (if/until regulators provide further clarification/guidance)

- An inclusive definition of "model" is expected when regulators determine the scope of AI/ML applications that fall under the guidance:
  - Definition of a model in SR 11-7/OCC 2011-12 includes statistical and/or mathematical techniques with inputs and outputs, which reflect characteristics of most AI/ML applications.
  - Historical experience suggests that regulators take a conservative view of what a model is (e.g., qualitative models based on expert judgment).
  - Supervisory matters regarding model risk management deficiencies remain outstanding at many firms.
  - Some regulators perceive that firms are inclined to underreport models.
- Banks should enhance existing model inventory management process to identify AI/ML models, which takes into account different AI/ML techniques, unique characteristics (e.g., dynamic calibration) and use cases/developers/sources/platform:
  - Banks should be prepared for an expansion of in-scope models given regulatory predisposition toward inclusion.
  - Models and other non-model components may be embedded in applications (e.g., robotic process automation (RPA) can coexist with ML); therefore, consistent model identification process is key to enable appropriate governance and control across all components.

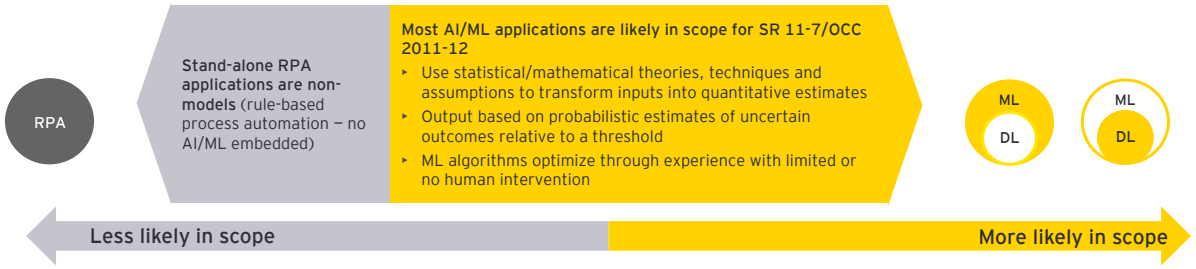


Figure 5: Model definition and related criteria should align with existing supervisory expectations

B. Risk assessment

It is also important for MRM functions to update their risk assessment frameworks to incorporate AI/ML attributes so that validators can adjust the scope and rigor of review to the risks posed by the model. Typically, banks assess a model's inherent risk based on its complexity, materiality and degree of reliance. Assessments of complexity should be enhanced by including large volumes of structured and unstructured data, frequency of retraining, opacity of the algorithm, number of hyper-parameters, reliance on open-source code and interrelationships with other risks. In addition, risk assessments should be updated to incorporate ethical and social implications as necessary, which will often require an evaluation of model explainability and the potential for bias and lack of fairness.

Current expectations also emphasize the importance of at minimum an annual review of models to determine whether they continue to be fit for purpose given the availability of new data or potential changes in the business, economic or regulatory environment. Annual review of AI/ML models to determine whether revalidation is necessary may be insufficient, given the dynamic nature of AI/ML models and the drift that can arise. The frequency of review should be predefined in the ongoing monitoring plan and linked to the risk assessment, with higher-risk models reviewed more frequently than lower-risk ones. This approach allows greater scrutiny over higher-risk models, which is consistent with current expectations that model assessments be risk-based.

C. Model development

As with traditional models, the development process for AI/ML models should cover the model's objective, selection process, design, initial parameterization, retraining, data sources and testing approach. Effectively deploying AI/ML requires the business problem statement to be clearly defined and the problem statement to be explicitly translated into:

- The transformation logic of the model (how the model transforms inputs into outputs)
- The selection of a model category (such as regression or classification)
- Model type (such as clustering, logistic regression, random forest, gradient boosting or neural network), including its specification

A clear definition helps developers evaluate whether the model works for its intended purpose, including its potential for harm. Consistent with traditional models, AI/ML models should also be developed in a well-controlled environment. For AI/ML, controls are designed to foster integrity, traceability and reproducibility of results. This requires the developer to keep a holistic perspective of all the elements of the life cycle — from data sourcing and pre-processing, model design and construction to implementation, performance assessment and ongoing monitoring — with controls embedded throughout.

Examples include controls regarding data sourcing, availability (in training and especially future production) and quality, manipulation and testing. Other examples include controls to achieve proper feature selection and engineering, calibration and training, as well as performance testing and monitoring. Users implement these controls in accordance with standards defined by the MRM function as it conducts its validation work.

As is done for traditional models, developers should perform appropriate tests and document the results. The goal of testing is to confirm that the model is working as intended. This entails assessing the accuracy of the model and its potential limitations, demonstrating its robustness and stability, and evaluating its behavior over a range of conditions (from base and stressed). The first step in testing is to define ex ante performance measures that are aligned with the goal embedded in the transformation logic and therefore the business purpose. Multiple measures should be employed to evaluate the accuracy of the model's representation.

The need for an AI/ML-specific technique should be justified given the incremental effort, and the ROI (return on investment) should be assessed with the investment in enhanced capabilities included. As with traditional models, developers should also thoroughly assess and document model choices, emphasizing the trade-offs and associated risks of alternative techniques.

Similarly, development requires choices regarding the algorithm design, the types of data, sources, definition of the goal, feature engineering, model selection and parameter optimization. These choices entail trade-offs, which are driven largely by the complexity of the problem statement and model. The impact of the trade-offs on model performance may be difficult to detect for some types of models given their inherent opacity (such as neural networks). These trade-offs include, but are not limited to:

- Predictive accuracy vs. interpretability of model output
- Algorithmic simplicity vs. computational intensity
- Bias vs. the variance of the model fit

Each of these choices also entails judgment. Firms should establish that the development of the more judgmental and qualitative aspects of their models is sound and well-documented. The judgmental nature of the choices and the risks associated with trade-offs reinforce the importance of subjecting trade-offs to testing.

Another area where AI/ML models can entail judgment is in the calibration of hyper-parameters.[8] These are typically chosen to reduce bias and variance, but depending on the methodology and data, computational feasibility could be affected. Developers should evaluate their hyper-parameter choices and assess their impact on model performance and results.

The level and effort of testing should be guided by the use case and an assessment of the inherent risk of the model. In preparation for testing, data sources should be explicitly identified, including the steps undertaken during pre-processing, feature engineering and testing to address the risks associated with data given its role as a potential driver of model risk.

Data quality issues or errors in sourcing, pre-processing and availability can undermine the quality of training data on an ongoing basis and therefore the effectiveness of commonly employed AI/ML techniques used to test model performance, such as cross-validation. These errors can lead to issues in data quality, bias, information security risks and/or privacy violations.

[8] Hyperparameters are numerical values that are set by the model developer prior to the estimation process to fit the model to the data. The model developer can select the parameter automatically using cross-validation or other techniques (e.g., Bayesian approach). Hyperparameters vary depending upon the specific AI/ML model. Examples include the number of layers or the width of each layer in a neural network model. Other examples are the number of and depth of trees in a random forest model or the penalty or regularization parameter used to control the number of variables in ridge regression to avoid overfitting.

D. Validation of conceptual soundness

We anticipate that regulators would view most AI/ML models as requiring independent validation and inclusion in MRM's model inventory given their historical posture toward the development of new quantitative approaches. The principles for validating traditional models under existing guidance are relevant for AI/ML models, even though the techniques may differ.

Like traditional models, AI/ML models should be subject to validation, according to their unique risks and intended business purpose, to affirm that they perform as expected, in line with their design objectives and intended business uses. The goal of validation is to assess key assumptions, limitations and potential impact to the firm. As a result AI/ML models should be evaluated for conceptual soundness and outcomes analysis and be subject to ongoing monitoring.

To remain consistent with current expectations, model validation frameworks and practices should be enhanced.

First, the validator should review the rationale for the use of an AI/ML model as opposed to more traditional techniques and whether the specification is informed by domain expertise and aligns with the business purpose. The validator should also confirm that the developer evaluated the risks associated with the trade-offs discussed above (see "Model development"). Validators should confirm that the model captures regulatory and ethical considerations where appropriate. As necessary, other control functions should be leveraged to verify that model specifications consider regulatory requirements.

To evaluate the conceptional soundness of an AI/ML model, validators should assess its design and construction, focusing on data integrity, feature engineering, hyper-parameter calibration, bias and explainability. The model's assumptions and the judgment used by developers for calibrating the model are also important to scrutinize.

1. Data integrity

AI/ML models rely on large volumes of heterogeneous and high-dimensional data, making it vital to document and trace lineage across the data life cycle — from sourcing to pre-processing and to training, testing and deployment — to establish that the data is appropriate and of high quality. A traceable data lineage increases the integrity of data (both the availability and quality) fed to the model and facilitates testing and validation.

The dynamic nature of AI/ML models and the need to manage feature changes throughout retraining also create challenges. As a result, validators must confirm that the data used in the model is of the same type, availability and quality that will be used in production. The degree to which validators can rely on the testing performed by developers should depend on the model's risk assessment. Validators should also test the robustness of different AI/ML techniques with respect to missing data, alternative normalization techniques, and anomalous or noisy data.

Firms may need to enhance their existing data remediation processes and associated testing infrastructure for model development and validation to address the high volume of structured and unstructured data that AI/ML models typically ingest. The enhancements could entail centralization of data and feature repositories to source, host, manage and govern data across AI/ML models to facilitate standardized remediation techniques (such as missing value fillers).

It is also important to enhance capabilities to manage the labeling process for supervised learning models and to assess how label errors can impact model predictions during development and retraining. Strategies could also include designing and building a standard development and testing environment that will also enhance standardization and techniques for easing model remediation.

Even when high-quality data is available, it may not be appropriate to use, given the concerns around privacy and/or information security. As a result, validators must confirm that developers are relying on data that is traceable, reliable and from approved sources. To this end, the validator should confirm that the sourcing and any pre-processing of the data were conducted in accordance with approved information security and privacy policies.

More generally, data testing first requires an effective data management framework that establishes a set of rules and standards on data quality, completeness and timeliness for AI/ML models, with considerations for data privacy, protection and ownership. The goal of the framework is to identify the risks associated with using data in ways that violate access and usage permissions articulated in policy. Without an effective framework, errors can result in the model inputs and labels, which can violate internal policies or regulations.

2. Feature engineering

Validators should also review feature engineering, a process in which input variables for the model are constructed from the raw data. Poor feature engineering triggers issues such as missing observations and artificial overlap between the target variable and features (i.e., leakage). These should be evaluated to avoid overfitting or underfitting in calibration.

Validators should also review the business intuition to select features, as well as any statistical analysis employed to reduce dimensionality (the number of attributes in a data set) and support the selection. Often, thoughtful business intuition and domain expertise can effectively reduce dimensionality. Statistical analysis is also typically employed to eliminate variables that are weakly correlated with the target variable. Examples include the Kolmogorov–Smirnov statistic and information value. Clustering analysis and dimensionality reduction (e.g., principal components) can be used to eliminate input variables that are redundant and do not add to the explanatory power or predictive accuracy of the model.

3. Sampling bias and fairness

When they assess conceptual soundness, validators should evaluate stakeholder impact, including bias and fairness, consistent with the use case and depending upon the model's inherent risk and complexity. Where necessary, validators should coordinate their evaluations with the other control functions (such as compliance for consumer applications).

In cases where other functions (e.g., compliance) have the requisite technical skill set to perform the assessment, validators can delegate the assessment to the other function and incorporate the results into the overall assessment of the model. Regardless of which function takes the lead in the assessment, close collaboration across functions is necessary to assess an AI/ML model for bias and fairness.

Sampling bias and fairness should be evaluated across the model's life cycle because it can arise in the design and construction of the model (e.g., the objective function and related transformation logic) as well as the input data and feature engineering. In general, bias occurs when the model improperly represents its target population. In practice, AI/ML models can exhibit many kinds of bias:

- **Sampling bias:** AI/ML models can exhibit sampling bias if they incorrectly and systematically underrepresent or overrepresent specific groups or classes from a population in a nonrandom fashion.[9] Data sampling processes should in principle generate a balanced training data set in which there are enough observations of the phenomenon of interest. In practice, obtaining balanced data sets is often difficult. Failure to evaluate sampling choices or exclusions may lead to the model being trained on a population that is too small or unrepresentative. Forms of bias include sample selection bias, statistical bias, survivorship bias, seasonality bias and omitted variable bias.
- **Fairness:** AI/ML models are largely based on pattern recognition and therefore lack commonsense reasoning regarding cause and effect. Thus, results may not make sense in the context of the business decision or, even worse, lead to a lack of fairness in results. This can arise if predictions are based on data that reflects institutional or societal bias (such as gender or race). Lack of fairness can also arise from sample selection bias or from how the objective function was defined. In consumer applications, the model result can lead to disparate treatment if there is implicit or explicit reference to group membership as a factor in the model or disparate impact if the outcome of the model on members of different groups varies.

To evaluate sampling bias, validators should assess the impact of data availability, representativeness, missing data, outliers, unbalanced samples, and the choice of imputation methodology on feature quality and bias. A key method for detecting sampling bias is to perform a deep conceptual review of the data processing steps (such as exclusions, vintages, sampling processes and reject inference) and target variable definition.

Improper sampling and associated bias can also arise from leakage. Validators should also assess the labeling process and the integrity of labels of target classes where the number of training samples is limited.

An assessment of fairness should begin with a clear and documented statement of the fairness principle, its relevance to the underlying business application and stakeholder impact. Validators should also consider whether a formal nondiscrimination criterion is necessary in the objective function and associated transformation logic. Several types of criteria can be used.[10] The appropriate criterion to select will depend upon how fairness is interpreted in the context of the business decision.[11]

As a result, individuals with domain expertise as well as other relevant control functions should be involved so the criterion aligns closely with regulatory requirements and fairness perceptions in the market or client base that the model is intended to serve. In the case of consumer models, validators should obtain explicit input and sign-off from the compliance function to confirm that consumer compliance risks have been appropriately addressed.

4 Hyper-parameter calibration

AI/ML model validation also needs to assess hyper-parameter calibration. The value of the parameter impacts the model's results and computational feasibility. Validators should evaluate how different parameter settings impact the model's results and the computational feasibility in production.

Stress testing and sensitivity of convergence and performance to changes in how hyper-parameters are set should be evaluated under different environments. Settings of hyper-parameters that led to a breakdown of the model should be identified. The choice of hyper-parameters should also be well-supported and documented. When changes are made, validators should confirm that the impact on model results is consistent with expectations.

[9] "Fairness: Types of Bias," *Machine Learning Crash Course website*, https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias. The term "sampling bias" is often used interchangeably with the term "selection bias" or "sample selection bias," although sampling bias can be considered as a special case of selection bias. Selection bias can be defined as bias arising from how a sample is chosen from a population and often refers to the negative impact on the validity of statistical tests arising from deficiencies in how the sample is chosen.

[10] See "Fairness and Machine Learning," Solon Barocas and Moritz Hardt and Arvind Narayanan, 2019.

[11] Ibid. There are no universally acceptable naming conventions for the types of bias or the criteria that can be selected. Examples include i) unawareness: sensitive attributes should be excluded; ii) demographic parity: the target variable should be independent of protected attributes, such as race, gender, etc.; the outcome of the model should be the same regardless of whether the protected attribute is used as an input; iii) equalized odds: the protected attribute is independent of other features, meaning that the result of the predictor conditioned on the outcome should not depend on the protected attribute; iv) predictive rate parity: the outcome and the protected attribute are independent of one another when each is conditioned on the predictor; and v) counterfactual fairness: understanding how the model performs when a different value is substituted for the value of the sensitive attribute. The impossibility theorem states that alternative fairness criteria cannot be satisfied at once.

5. Explainability

Explainability entails understanding how a model produces outputs based on the input variables and being able to interpret the outputs in a qualitative fashion.[12] For most traditional models, the model's design can facilitate explainability. For example, variables in linear regression can be aligned very closely to factors that are derived from domain expertise. Sensitivity analysis and stress testing can be used to determine which variables contribute to the model's output. The results of the analysis and testing can be readily compared to a user's business intuition and domain expertise to determine whether the results are reliable.

For certain types of complex AI/ML models, such as neural networks and ensemble techniques, the way outputs respond to inputs may be unclear without enough transparency, which reduces a user's confidence in the model's reliability and its results. Some AI/ML models may not be traceable, meaning it is difficult to understand how inputs get transformed into outputs, or explainable on a stand-alone basis (meaning that their outputs cannot be attributed to the variables driving them without using additional techniques). A lack of transparency can undermine an assessment of conceptual soundness, especially as related to sampling bias and fairness, because it makes it difficult to understand whether models are successfully meeting testing objectives and are fit for purpose.

In August 2019, the Bank of England published a paper that provides a useful framework for assessing a model's explainability, based on five key questions:[13]

- Which features mattered in individual predictions?
- What drives the actual projections more generally?
- What are the differences between an ML model and a linear one?
- How does the ML model work?
- How will the model perform under the new states of the world (that aren't captured in the training data)?

To address these questions, it is essential to conduct stress tests and sensitivity analyses, which are important aspects of current expectations for traditional models and evaluating the risks of AI/ML. Different approaches are available to implement a framework for explainability.[14]

One approach is to assess the importance of input features to the model predictions. Importance can be evaluated "globally," where the overall impact of an input feature on model predictions is assessed. Examples include tree-based importance, permutation test-based importance and global sensitivity analysis. Importance can also be evaluated "locally," where the effect of an individual observation's attributes on the model's prediction can be evaluated. Examples include Shapley values, such as LIME and its variants, DeepLIFT and layer-wise relevant propagation (LRP).

Importance assessments should be performed as part of the validation. Understanding local importance is particularly relevant in assessing consumer models of AI/ML. In addition, to identify the direct relationship between model inputs and predictions, professionals can use approaches such as one-dimensional and two-dimensional partial dependence plots (PDP), individual conditional expectations (ICE), interaction statistics (IS), marginal plots and accumulated total derivative (ATDEV) plots.

Surrogate models can also be employed, which entails using a simpler and more transparent model (e.g., tree, linear regression) as a proxy for a less transparent model (e.g., neural network). Global surrogates (e.g., global surrogate tree) use the entire data set to proxy for the original model. Local surrogates (e.g., LIME and variants) use subsets of the data to proxy.

Mitigating controls can also be recommended and adopted after validation to address a lack of transparency in how inputs impact model outcomes. The type of the mitigating controls can vary based on the use case and risk assessment of the AI/ML model. For example, controls can be defined for a range or acceptance criteria for output, whereby the model output is used only if it is within the defined range of pre-specified criteria. Exception procedures can be defined for out-of-range output that would require review by a human operator.

Controls may also include more stringent or more frequent ongoing monitoring. This would entail ongoing assessment and, where necessary, testing of input data to identify outliers or cases different from the data on which a model was trained. It could also entail the use of benchmark models to compare outputs and variances against predefined thresholds to trigger further investigation, revalidation or use of alternative models.

---

[12] A term used interchangeably with explainability is interpretability, but some authors cite differences. A model is interpretable when it can be understood by human beings. A model is explainable when it provides a rationale for its results. See "Explaining Explanations: An Overview of Interpretability in Machine Learning," Leilani H. Gilpen, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal, 2019.

[13] "Machine learning explainability in finance: an application to default risk analysis," *Bank of England website*, https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis, 9 August 2019.

[14] Sensitivity analysis and stress testing are necessary to meet the explainability criterion as we define it, but they may not be sufficient to engender trust depending upon the complexity of the model, which is why other aspects of the MRM framework, including mitigating controls, are important. For example, some authors emphasize that deep neural networks can only foster trust and achieve explainability when they are able to provide reasons for their decisions based on cause and effect and counterfactual analysis. Ibid Gilpen et. al.

E. Outcomes analysis

In consultation with other control functions, validators should review the model's performance metrics — both statistical and computational ones — as necessary to determine whether the number of metrics and the results support a conclusion of whether the model is appropriate for its intended purpose.

Results of the performance metrics should inform the limitations imposed on the models. Many metrics can be defined for AI/ML models, depending upon whether the ML problem is supervised or unsupervised. In the case of supervised learning, the appropriate measure depends upon whether the model is a regression or a classification problem.

In the case of unsupervised learning, where there is no target variable, metrics can be defined and evaluated based on whether observations within a particular class are close to one another and observations of different classes are not based on a distance measure (e.g., Euclidean distance) used to define the goodness-of-fit.[15] In the case of supervised learning, where the target variable is fitted to a set of features, goodness-of-fit measures depend on the model type (such as random forest or neural network) and include the mean-squared-error (MSE), Gini coefficient, entropy and more. MSE, which is among the most common, can be decomposed into bias and variance, and explicit evaluation of the trade-offs (see "Model development") should be reviewed and documented.

For classification problems, where the target variable takes a discrete value (e.g., "0" or "1" that corresponds to a state or condition), the "confusion matrix" can be used for outcomes analysis. The objective is to determine how well the model classifies an observation based on a series of input features.

Performance metrics include accuracy, precision, recall and the F-Score.[16] Precision refers to false positives (Type 1 errors), and recall refers to false negatives (Type 2 errors).

Thresholds for an acceptable level of false positives or negatives should be specified ex ante in development and reviewed by validators to confirm consistency with the business problem. Performance should be evaluated against the threshold. These thresholds can be evaluated using a receiver operating characteristic (ROC) curve, which depicts the trade-off between precision and recall and therefore allows the separation of positive and negative values. The area under the curve (AUC) is the area under the ROC with a value close to 1 showing high separability between the two cases.

As noted above, performance results should be evaluated for sampling bias and fairness. Pay close attention to models with very high accuracy and recall that such accuracy may overstate model performance when there are too few observations of the phenomenon of interest.

Out-of-sample testing is also an important component for outcomes analysis for AI/ML. Out-of-sample testing can be evaluated using cross-validation in conjunction with ensemble learning techniques.[17] The predictions of the model with the best fit on the training data set are used to compare with the validation set and then ultimately the test set. Learning curves can be constructed from the model's training and test set errors to help understand the degree of statistical bias and variance of the model and to evaluate the trade-offs between the two.

[15] The goal of unsupervised learning is typically to identify structure in the data by, for example, grouping observations into categories. Examples include k-means clustering, hierarchical clustering and recommender systems. In general, these approaches assign observations to categories based on the measured distances between an observation assigned to a category and the "center" of the observations assigned to that category

[16] The F-Score is a blend of the precision and recall into a single statistic (i.e., the harmonic mean).

[17] In the case of k-fold cross-validation, for example, which is among the most common techniques, k subsets of the data are withheld from the training set and used to test the model's performance.

F. Use

Because they are in a position to observe performance, users, which can include developers and owners, should play a role in evaluating model performance over time through an established feedback mechanism. This creates checkpoints for user intervention over the AI/ML model life cycle and gives users an opportunity to effectively challenge model results.

Users require effective and fit-for-purpose models, and, as necessary, they should consult with other control functions (such as compliance and operational risk management) about model performance concerns.

Checkpoints for users are important because models can appear to perform well, but sometimes the performance only holds over a narrow set of conditions or thanks to factors unrelated to the features of the model. Models can also be subject to "drift" over time, which can go undetected unless users are involved.

For their intervention to be effective, model performance needs to be explainable. Users must understand the sensitivity of performance to changes in inputs at inception and over time to determine whether performance is consistent with their domain expertise and intuition.

When certain features unexpectedly explain or fail to explain results over a range of conditions, the user should see this as an indication that the model may not be performing as intended. The user should then contact developers and validators about potential performance issues. In addition to helping detect poor model performance, users can determine whether the model is remaining true to the original business purpose and achieving the desired business outcome.

G. Change management

Developers should confirm that the systems infrastructure can support the performance requirements of the model with respect to data capacity, retraining and calibration. They should also affirm that the model has been configured and integrated properly into the production environment. Errors can arise when firms employ legacy systems, upgrade from one model version to another, or migrate the model from one programming environment into another.[18]

When discrepancies between the testing environment and the production environment arise, model performance can be undermined and testing results invalidated. To address these risks, developers and users should confirm that minimum standards for deployment have been met.

Developers should also assess the computational feasibility of the model in the production environment. Testing should verify that the optimization algorithm that typically underlies the transformation logic is converging properly and generating sensible results, as well as confirm that the model performs over a range of "call conditions."[19] Stress testing the model under different conditions would be important to understand the model's stability and robustness in production.

Active changes must be monitored in input data against the training data to confirm data quality and the statistical consistency of the new data with the training data going forward. This validates that the data-generating process is the same. Changes in the input data could also require changes in the production environment.

It is also important that developers create a comprehensive ongoing monitoring plan to confirm that the model is operating as intended over time. The plan should consider model performance (e.g., drift), stability and alignment with business purpose. The plan should rely on the performance indicators and thresholds established in development to determine the degree of performance deterioration that would warrant further review or revalidation. The performance indicators should be evaluated after the model is retrained to insignificant changes in feature importance.

Real-time circuit breakers to set up performance boundaries for AI/ML models can also be an effective tool to establish that models are performing as intended. When performance boundaries are breached, benchmark or legacy models can be pre-specified and employed as fallback options. The monitoring plan should include checks to confirm the processing power for the model remains adequate so that the model can be available and reliably accommodate potential usage increases.

[18] Porting can be necessary when AI/ML models are prototyped and tested in one language and translated into another to align with the existing infrastructure. In other cases, models can be "ported" using "containers" (e.g., docker, kubernetics) and related tools in conjunction with an API-based architecture (e.g., microservices) to facilitate integration.

[19] A "call" refers to the number of times users access a model when in production.

H. Ongoing monitoring of AI/ML models

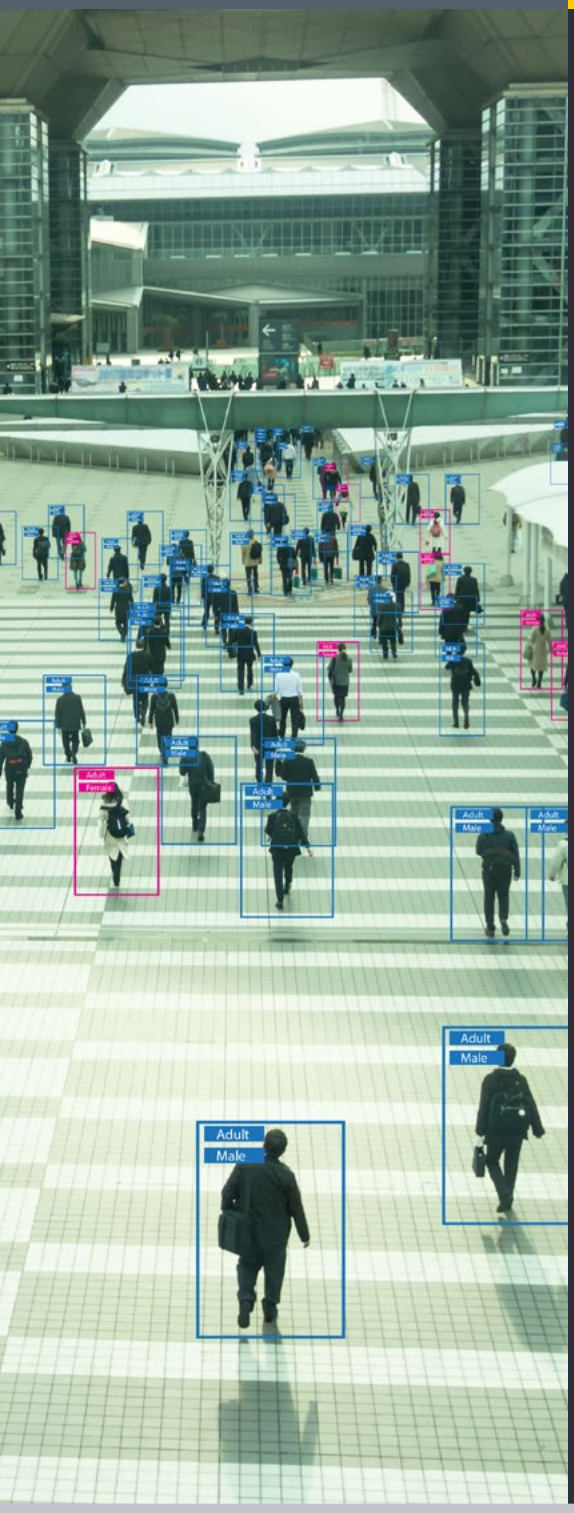
Validators should review all ongoing monitoring plans and consult with other control functions as necessary to verify they are appropriate, given the inherent risks of the model. Validators should confirm that the plan aligns with the risk assessment, considers model performance (e.g., drift), stability and alignment with business purpose, as well as that the performance indicators selected for the plan are appropriate, given the intended business purpose. Finally, validators should confirm that the performance indicators are monitored at an appropriate frequency, given how frequently the model is retrained.

Standards should differentiate between passive and active changes. Ongoing monitoring can be challenging for AI/ML models because dynamic retraining makes it difficult to define what constitutes a model change and how to assess it. Frequent retraining can lead to passive changes, even when it is in accordance with a documented and approved retraining approach. Passive changes can lead to changes in the feature importance of the model, which could be tantamount to a model change.

As with traditional models, active changes can also be made to model methodology, input types, use, monitoring approach and more, which can be considered as a model change. In the case of active changes, it is important to evaluate the change in input data against the training data to confirm data quality and the consistency of the new data with the training data going forward.

4

Address third-party and open-source considerations in validation



Many AI/ML models are directly obtained from open sources or are developed by third parties (in some cases leveraging open sources). In either case, these models should be subject to the enhanced standards for MRM described previously as well as other risk and control frameworks (e.g., privacy, information security) as appropriate.

When a firm relies on third-party capabilities for development, a risk assessment of the model should inform the expectations for validation and testing. At minimum, the third party should demonstrate a rigorous development and testing process that addresses the attributes of AI/ML models.

Testing results should be requested and made available where the testing is relevant to the business purpose and portfolio composition. Additional testing, such as benchmarking and sensitivity analysis, should be required as per the risk assessment to compensate for model opacity and lack of explainability. The third party should also describe its approach to ongoing monitoring, which should employ the enhanced capabilities described above, and outcomes analysis. In addition, firms should evaluate the vendor's data risk management practices to confirm that the vendor tests for bias and fairness.

5

Enhance governance, policies and controls

The oversight of AI/ML models should be consistent with the processes used for traditional models. Board and senior management oversight remains important. They should be aware of use cases being employed and understand the effectiveness of governance and controls used in the AI/ML model life cycle. Roles and responsibilities for model developers, users and validators, and other control functions should be clearly articulated to achieve ownership and accountability for risks. Internal audit will also need to remain engaged to give assurance that the MRM framework and related controls are effective for AI/ML models.

Nevertheless, several enhancements to policies and procedures should consider the dynamic and integrated risks associated with AI/ML. MRM policies should explicitly reference how other risk and control requirements (e.g., information security) apply where appropriate. That way, AI/ML model developers have clarity on all requirements needed to get models approved and control functions understand how their responsibilities are allocated. Procedures associated with enhanced capabilities and their relationships to other policies should also be well documented.

Contacts:

Gagan Agarwala
Principal, Ernst & Young LLP
gagan.agarwala@ey.com

Alejandro Latorre
Principal, Ernst & Young LLP
alejandro.latorre@ey.com

Susan Raffel
Partner, Ernst & Young LLP
susan.raffel@ey.com

Rushabh Mehta
Principal, E&Y Advisory Services Ltd
rushabh.mehta@hk.ey.com

Jan Zhao
Principal, Ernst & Young LLP
xiaojian.zhao@ey.com

Anvar Nurullayev
Senior Manager, Ernst & Young LLP
anvar.nurullayev2@ey.com

Brian Clark
Senior Manager, Ernst & Young LLP
brian.clark@ey.com

Rui Tang
Senior Manager, Ernst & Young LLP
rui.tang@ey.com

EY | Assurance | Tax | Transactions | Advisory

About EY
EY is a global leader in assurance, tax, transaction and advisory services. The insights and quality services we deliver help build trust and confidence in the capital markets and in economies the world over. We develop outstanding leaders who team to deliver on our promises to all of our stakeholders. In so doing, we play a critical role in building a better working world for our people, for our clients and for our communities.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. For more information about our organization, please visit ey.com.

Ernst & Young LLP is a client-serving member firm of Ernst & Young Global Limited operating in the US.

US SCORE no. 09129-201US
2002-3400789 BDFSO
ED None

ey.com