Model Risk Management AI and ML Roundtable

Current thinking and risks identified in artificial intelligence and machine learning (AI and ML) adoption

October 2025

Agenda

1. Background and introductory remarks (10 min)
2. Presentation: Current thinking and key risks identified in AI and ML adoption (20 min)
3. Discussion: Challenges and what guidance is needed to further support AI and ML adoption (80 min)
4. Closing remarks (10 min)

Introduction: Model Risk Management, AI and ML

- The PRA continues to treat model risk management (MRM) as a **strategic supervisory focus area**[1].
- The publication of Supervisory Statement (SS) 1/23 '*Model Risk Management principles for banks*' sets out principles-based expectations to support firms in developing effective MRM frameworks for all model types, including models that use Artificial Intelligence (AI) and Machine Learning (ML) technologies.

The PRA **continues to engage with firms** to advance the understanding, identification and management of model risk, including AI and ML, via the following initiatives:

- MRM roundtables to discuss thematic findings and concerns.
- The AI Consortium[2] (AIC),which provides a platform for public-private engagement to further dialogue on the capabilities, development, deployment, use, and potential risks of AI in UK financial services.

Our aim today is to share our current thinking on risks identified in AI and ML adoption, in the context of implementing the expectations set out in SS1/23.

We would also like to hear your views on current challenges to understand whether frameworks under SS1/23 are sufficient to mitigate these risks and address these challenges.

[1] Letter from Charlotte Gerken and Laura Wallis 'UK Deposit Takers Supervision: 2025 priorities'
[2] Artificial Intelligence Consortium | Bank of England

Current thinking and risks identified in AI and ML adoption

MRM AI and ML Governance Review 2025

1. Risk appetite

- The board should set a model risk appetite that **clearly articulates the level and types of model risk** the firm is willing to accept. (Principle 2.1c) of SS1/23).
- Due to their opaque nature and the lack of transparency compared to "traditional" models, AI and ML models introduce higher uncertainty.
- Establishing and expressing risk appetite before deploying AI and ML models could help to reduce the risk of deploying models that exceed firms' own tolerance for risk.
- Evolving governance frameworks that address the complex and emerging risks associated with AI and ML models could support consistency of MRM across the model lifecycle. For example, triggers for model re-validation that are explicitly linked to the risk appetite.

2. Model Tiering

Observation

- Instances were observed where firms' model tiering policies were not aligned with their model inventory submissions.

  For example, a policy may state that AI and ML models will have a minimum categorisation of medium complexity, but in the corresponding inventory, AI and ML models were classified as low complexity.

- Inaccurate / absent information in model inventories lead to incomplete view of aggregate model risk, underdeveloped model risk appetite and a failure to adhere to the prescribed approach to risk management.
- While individual models may have low materiality or complexity when assessed in isolation, deploying identical AI and ML techniques across multiple jurisdictions or portfolios can result in a higher aggregate complexity or materiality. This reinforces the need for adequate challenge to the tiering approaches.

3. Explainability and Interpretability (1/2)

Explainability and Interpretability (E/I) **are key factors** when determining a model's complexity under SS1/23.

E/I techniques such as SHAP, LIME, and Partial Dependence Plots (PDPs) are now commonly used to interpret AI and ML models by attributing the contribution of each input feature to the model's predictions.

However, our research highlights a critical limitation: E/I techniques rely on the assumption that **input features are independent** (or at least uncorrelated). **This assumption rarely holds** in real-world big data, because there is usually a certain degree of correlation amongst features.

When features are correlated, E/I techniques may distort feature importance rankings. The most important features fluctuate in rank, while less relevant variables can be incorrectly prioritised.

Explainability and interpretability in Machine Learning: do popular methods deliver on their promises?
Ivona Cickovic and Andrea Serafino (Quantitative Modelling and Analysis, MDRD, SRS)
Forthcoming research, expected publication in *Bank Underground*, January 2026

3. Explainability and Interpretability (2/2)

True data		Feature importance								PDP								Shapley								LIME							
Abs beta coeff-based rank order		LM		RF		GBM		NN		LM		RF		GBM		NN		LM		RF		GBM		NN		LM		RF		GBM		NN	
X50	1	X7	2	X31	4	X50	1	X31	4	X7	2	X15	10	X50	1	X31	4	X50	1	X31	4	X50	1	X31	4	X7	2	X31	4	X50	1	X31	4
X7	2	X50	1	X15	10	X40	35	X50	1	X50	1	X31	4	X15	10	X7	2	X7	2	X15	10	X40	35	X43	19	X50	1	X15	10	X15	10	X17	36
X16	3	X16	3	X40	35	X15	10	X43	19	X16	3	X40	35	X40	35	X12	6	X12	6	X40	35	X14	8	X24	24	X12	6	X40	35	X40	35	X7	2
X31	4	X12	6	X5	9	X31	4	X24	24	X12	6	X5	9	X14	8	X24	24	X16	3	X14	8	X31	4	X50	1	X16	3	X5	9	X14	8	X41	26
X44	5	X14	8	X14	8	X14	8	X17	36	X14	8	X14	8	X31	4	X50	1	X14	8	X5	9	X15	10	X41	26	X14	8	X14	8	X31	4	X50	1
X12	6	X31	4	X25	22	X25	22	X7	2	X44	5	X25	22	X25	22	X17	36	X31	4	X50	1	X25	22	X17	36	X44	5	X43	19	X24	24	X3	46
X37	7	X44	5	X50	1	X24	24	X41	26	X31	4	X50	1	X24	24	X3	46	X37	7	X43	19	X24	24	X12	6	X37	7	X50	1	X25	22	X43	19
X14	8	X5	9	X43	19	X41	26	X12	6	X5	9	X43	19	X41	26	X43	19	X5	9	X25	22	X41	26	X30	20	X31	4	X25	22	X41	26	X12	6
X5	9	X37	7	X41	26	X12	6	X30	20	X37	7	X41	26	X7	2	X41	26	X44	5	X41	26	X12	6	X7	2	X8	14	X41	26	X12	6	X30	20
X15	10	X8	14	X24	24	X7	2	X3	46	X11	11	X7	2	X12	6	X30	20	X8	14	X24	24	X8	14	X3	46	X11	11	X24	24	X8	14	X24	24

Key message: Using E/I techniques on an AI / ML model is likely to mislead regarding importance of different features and therefore mislead regarding model behaviour and interpretation.

4. Data and Overfitting

Concerns

AI and ML models are particularly prone to overfitting due to their high parameter count and sensitivity to noise. This vulnerability is amplified **when datasets used in model development fail to adequately represent the target population**, reducing the model's ability to generalise to real world conditions.

What is the risk?

AI and ML models may exhibit significant performance deterioration when applied to instances that differ from the datasets used to develop the model.

Comments

Give careful consideration to datasets used for developing AI and ML models.
Large and heterogeneous datasets often help reduce overfitting.

Standard validation splits may not detect generalisation issues in non-representative datasets since both the training and testing sets originate from a specific data sample that may not be representative of the target population.

5. Model Development Testing / Independent Validation

While statistical techniques such as cross-validation and performance metrics remain foundational, **their underlying assumptions** (such as independence and stationarity) often do not hold in AI / ML or big data contexts.

Well-researched testing criteria that suit the specific nature of the data and modelling context could improve reliability compared to **routine** testing procedures.

Furthermore, models are frequently tested on clean, well-curated datasets, yet deployed in noisy, adversarial, or dynamic environments. This disconnect can result in testing outcomes that are overly optimistic and not reflective of real-world performance.

6. Other observations
(Principle 3: Model development, implementation and use)

Model development testing should be conducted when either:

- Material model changes are made.
- Cumulative material changes occur over a period of time for dynamic models. (Principle 3.3c) of SS1/23)

  Tracking cumulative non-material changes is a practical way to prevent model behaviour or risk profiles from drifting without appropriate oversight.

Model selection:

- It is important to assess the trade-offs between model performance, complexity, explainability, and reliability. In some cases, **simpler or more established/traditional techniques may offer more appropriate or reliable outcomes**, particularly where performance gains from AI and ML models are limited.

7. Ongoing Model Monitoring - Frequency (1/2)

What we have observed

Firms often stated that ongoing model performance monitoring periodicity is tier-dependent but broadly did not prescribe **quantitative expectations/limits** within policies.

What is the risk?

AI and ML models, particularly models that can dynamically recalibrate, can quickly and cumulatively evolve beyond their validated risk appetite.

Comments

The current periodicity of ongoing monitoring for AI and ML models is **not frequent enough**.

For example, six-month intervals may be insufficient to determine whether a dynamic AI and ML model is performing as expected and remaining within its risk appetite. Firms may need to consider whether the current **frequency** and **scope** of ongoing model performance monitoring remain appropriate for AI and ML models.

7. Ongoing Model Monitoring – Performance Degradation (2/2)

Given the risk of rapid performance degradation in AI and ML models, firms should consider implementing robust monitoring frameworks that **identify early signs of deterioration** and ensure alignment with intended use and risk appetite.

It is possible that AI and ML models deteriorate so abruptly that it may leave limited time for remediation. To manage this risk, firms should consider developing **pre-approve fallback models or** challenger models and, in extremis, kill switches.

Performance degradation is a consequence of poor model development testing. To help reduce the risks of degradation, firms could also **define clear, quantitative, and measurable triggers** that prompt model re-validation procedures when degradation is detected.

Discussion

What are your current challenges with AI and ML adoption, are existing MRM policies sufficient to mitigate the risks posed, and what further guidance is needed from us for AI and ML?

Closing remarks

Thank you for attending