ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING: A MODEL RISK MANAGEMENT PERSPECTIVE

AUGUST 2022

Author: **Alexey Rubtsov,**
PhD, Senior Research Associate, Global Risk Institute

GLOBAL RISK INSTITUTE

INTRODUCTION

The last decade has witnessed a large-scale adoption of Artificial Intelligence and Machine Learning (AI/ML) models in finance. Although there are many benefits that AI/ML can bring to financial services (e.g., higher accuracy, automation), it could also introduce new and amplify existing risks. In this respect, financial regulators around the world are currently working on regulatory requirements that AI/ML models should satisfy when applied by financial institutions. In this report we discuss some of the most recent developments in AI/ML model risk management.

Important characteristics of AI/ML models that make it necessary for regulators to update existing guidelines are:

- Complexity. Released by Meta in July 2022, No-Language-Left-Behind (NLLB-200) language translation AI model includes 54 billion parameters!
- Explainability. Many AI/ML models are black boxes in the sense that it is difficult to understand models' behavior, explain how they arrive at their conclusions, and, therefore, manage model risk.
- Ability to learn in real time. Models that apply Reinforcement Learning can adjust their behavior when new data becomes available. This feature makes it more difficult to foresee where the model could go wrong.
- Use of unstructured data. AI/ML models can be trained on texts, audio, and video data which are more challenging to verify for quality and completeness.

As a result of the above characteristics of AI/ML models, Model Risk Management (MRM) practices need to evolve in order to realize the full potential of AI/ML while minimizing the unintended negative impacts. In this paper we discuss some recent efforts to manage risks of models used by financial services industry in the following jurisdictions: the European Union, the United Kingdom, Singapore, United States, and Canada.

THE EUROPEAN UNION

In April 2021 the European Union issued the Artificial Intelligence Act, draft rules that apply to all industries (except military) including financial services.[1] The AI Act combines a risk-based approach with a layered enforcement mechanism: a lighter legal requirements applies to AI applications with a minimal risk, and applications with an unacceptable risk are banned. The risks of use cases are split into four categories: minimal (e.g., spam filters), limited (e.g., chatbots), high (e.g., credit decisions), and unacceptable (e.g., social scoring). It may take another year before the AI Act becomes a legally enforced law.

The definition of AI used in the draft rules is very broad and includes rather standard tools commonly used in modelling (e.g., Bayesian estimation). However, it does not necessarily mean that such tools will be subject to the proposed regulatory rules as it will depend on the risk/materiality of the use case.

1 See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

THE UNITED KINGDOM

In this section we discuss two recent documents: the report from the AI Public-Private forum and the consultation paper on Model Risk Management Principles for Banks.

AI Public-Private Forum

In February 2022, the Bank of England and the Financial Conduct Authority jointly released their final report based on the AI Public-Private Forum whose goal is to further the dialogue between the public sector, the private sector, and academia on AI.[2] The report covers the entire financial sector of UK. Risks from AI were considered as stemming from three areas: Data, Model, and Governance. Below we list the key findings of the report.

Data. Given the complexity of data that could be used by AI models, data attributes such as provenance, completeness, and representativeness are identified as the most critical markers to understand. Since the data can change, firms should pay attention to documentation, versioning, and ongoing monitoring of their datasets. Special attention should be paid to data sourced from third parties.

Model. Model complexity and explainability are identified as the most challenging issues to address. To ensure models behave as expected, monitoring their behaviors is key.

Governance. Capacity of AI models for autonomous decision-making is a big challenge to ensure accountability and responsibility within financial institutions. Since AI models will invariably interact with existing risk and governance processes, such frameworks as data governance, MRM, and operational risk are suggested to be a good starting point for establishing an AI governance framework. In addition, the following suggestions were made:

- governance of AI systems should reflect the risk and materiality of the use case.
- a centralized body within firms should set the AI governance standards.
- diversity of skills and perspectives are important to help manage the complexity of AI systems.
- ensuring that there is an appropriate level of awareness of AI's benefits and risks throughout the organization.

Model Risk Management Principles for Banks

In June 2022, the Bank of England issued a consultation paper that sets out the Prudential Regulation Authority's (PRA) proposed expectations regarding banks' management of model risk.[3] One of the reasons for the consultation paper is the increasing use of more complex models including AI/ML models. The proposed expectations would not apply to third-country firms operating in the UK through a branch. The PRA proposes all firms adopt the following five principles which it considers key in establishing an effective model risk management (MRM) framework.

1. **Model identification and model risk classification.** Firms should have a clear definition of what is counted as a "model" and a risk-based tiering approach to categorize models.
2. **Governance.** Organizations should establish strong governance oversight with a board that sets clear model risk appetite.
3. **Model development, implementation and use.** This principle implies that firms should have a robust model development process with standards for model design, implementation, model selection, and model performance measurement. Testing of data, model construct, assumptions, and model outcomes should be performed on a regular basis.

2 See https://www.bankofengland.co.uk/research/fintech/ai-public-private-forum

3 See https://www.bankofengland.co.uk/prudential-regulation/publication/2022/june/model-risk-management-principles-for-banks

4. **Independent model validation.** A validation process that provides ongoing, independent, and thorough testing to model development and model use should be present within a firm.

5. **Model risk mitigants.** Organizations should establish policies and procedures for the use of model risk mitigants when models are not performing well and have procedures for the independent review of subsequent adjustments.

Note: The above principles are similar to the Guidance on Model Risk Management (also known as SR11-7) which is applied to banking organizations in the United States. This similarity helps promote consistency across jurisdictions. It is also worth noting that the proposed principles are intended to complement, not supersede existing requirements that are currently in force for certain model types (e.g., credit risk, counterparty credit risk).

SINGAPORE

Veritas consortium, comprising the Monetary Authority of Singapore (MAS) and various industry partners, launched an initiative that aimed to enable financial institutions to evaluate their AI and Data Analytics (AIDA)-driven solutions against the set of 14 principles that promote Fairness, Ethics, Accountability and Transparency (FEAT) in the use of AIDA by the financial sector.[4]

Fairness. There are four Fairness principles that focus on Justifiability (1 and 2) and Accuracy/Bias (3 and 4) of AIDA-driven decisions.

1. Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified.

2. Use of personal attributes as input factors for AIDA-driven decisions is justified.

3. Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimize unintentional bias.

4. AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.

Note: Justifiability addresses the issues around the use of sensitive information: if a firm can justify the use of a particular factor (e.g., age), the use can be considered aligned with the first two principles of Fairness. It is also emphasized that the frequency of revisions in principles one and two depends on the materiality of the use case.

Ethics. The following two principles are suggested to adhere to Ethical standards.

1. Use of AIDA is aligned with the firm's ethical standards, values and codes of conduct.

2. AIDA-driven decisions are held to at least the same ethical standards as human-driven decisions.

Accountability. Accountability principles focus on Internal Accountability (1, 2, and 3) and External Accountability (4 and 5).

1. Use of AIDA in AIDA-driven decision-making is approved by an appropriate internal authority.

2. Firms using AIDA are accountable for both internally developed and externally sourced AIDA models.

3. Firms using AIDA proactively raise management and Board awareness of their use of AIDA.

4. Data subjects are provided with channels to enquire about, submit appeals for, and request reviews of AIDA-driven decisions that affect them.

5. Verified and relevant supplementary data provided by the data subjects are taken into account when performing a review of AIDA-driven decisions.

Transparency. There are three principles that should ensure adherence to the Transparency requirement.

1. To increase public confidence, use of AIDA is proactively disclosed to data subjects as part of general communication.

4 See https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions

2. Data subjects are provided, upon request, clear explanations on what data is used to make AIDA-driven decisions about the data subject and how the data affects the decision.

3. Data subjects are provided, upon request, clear explanations on the consequences that AIDA-driven decisions may have on them.

Note: Increased transparency could increase the risk of exploiting and manipulation of AIDA models. In addition, explanation should not lead to the exposing of intellectual property or publishing proprietary source codes.

To translate these principles into practical implementation by financial institutions, a complete set of five white papers describing the corresponding methodologies were published by MAS in February 2022.

THE UNITED STATES

In February 2022, U.S. Democratic lawmakers introduced a bill in both the Senate and the House of Representatives titled the "Algorithmic Accountability Act of 2022".[5] This bill aims to bring more transparency and oversight of models that are used to make automated decisions.

According to the bill, companies that use or supply algorithmic tools will have to conduct assessments of the tools if they expect them to be used for making critical decisions.

Annual reports about the assessments will have to be submitted to the Federal Trade Commission (FTC) which will be making the rules for algorithmic impact assessments if the bill passes. The FTC would publish a repository, available to the public, that would contain information about the automated systems based on the reports. As for what constitutes a "critical decision" the act lists a few categories that have a substantial impact on individuals' lives such as access to, or the cost of, education, employment, essential utilities, family planning, financial services, healthcare, housing, and legal services.

CANADA

In September 2020 the Office of the Superintendent of Financial Institutions (OSFI) published a discussion paper which identified three core principles to manage risks associated with the use of AI/ML by financial institutions.[6] The goal of the paper was to seek feedback on advanced analytics tools (including AI/ML) in order to enhance existing MRM guidelines to accommodate AI/ML models.[7] An industry letter "Proposed Revisions to Guideline E-23 on Model Risk Management" was published in May 2022.[8] The suggested principles for MRM are Soundness, Explainability, and Accountability.

Soundness. Soundness is a broad and complex principle that takes into consideration, among other things, issues pertaining to data, model development, validation, monitoring, bias, and documentation.

Explainability. Explainability addresses the requirement to understand and describe the mechanics of the model and meaningfully explain the results. Degree of model explainability is suggested to depend on materiality of the use case, among other factors.

Accountability. Firm's risk management frameworks should integrate AI/ML models and clear roles and responsibilities should be assigned across the institution.

OSFI is set to release a draft comprehensive guideline in March 2023 for public consultation slated for release in the fall of 2023.

In June 2022 Bill C-27 was introduced in Parliament as an attempt at reforming federal privacy law in Canada. Part 3 of the bill is devoted to the Artificial Intelligence and Data Act

5 See https://www.congress.gov/bill/117th-congress/house-bill/6580/text

6 See https://www.osfi-bsif.gc.ca/Eng/fi-if/in-ai/Pages/tchrsk-sm.aspx

7 The current guidelines are Guideline B-9 (Earthquake Exposure Sound Practices), Guideline E-23 (Enterprise-Wide Model Risk Management for Deposit-Taking Institutions), and Guideline E-25 (Internal Model Oversight Framework), which pertains to property and casualty insurers.

8 See https://www.osfi-bsif.gc.ca/Eng/fi-if/in-ai/Pages/E-23_let.aspx

(AIDA) which aims to regulate the development and use of AI in the private sector. AIDA sets out specific requirements for so-called "high-impact systems" which is similar to the risk-based approach of the EU.

AIDA proposes new governance and transparency requirements for businesses that use, develop, and design AI. Such businesses must establish measures to identify, assess and mitigate the risks of harm or biased output. In addition, information about intended and actual uses must be made publicly available. The bill also focuses on processing or making available for use any data related to human activities for the purpose of designing, developing or using AI.

SUMMARY AND CONCLUSIONS

The following table provides the summary of discussed approaches across jurisdictions.

There are a few characteristics where the discussed approaches align.

- AI/ML models should be conceptually sound (e.g., accurate, reliable, robust, sustainable); explainability is essential in high-stake decisions
- Organizations should have proper governance structures that address challenges created by AI/ML models (e.g., transparency, accountability)
- AI/ML models should not cause any harm to individuals and society (e.g., bias, discrimination, ethical considerations, privacy concerns)

Although the approaches proposed in different jurisdictions are not legally enforceable laws, the issues covered by the referenced reports provide a strong basis for regulators to write future-proof guidelines that would welcome innovation and ensure that the stability of our financial system would not be undermined.

Jurisdiction	Document	Proposed approach
EU	The AI Act	Risk-based: limited, minimal, high, and unacceptable risks
UK	Report from AI Private-Public forum	Focus on: Data, Model, Governance
	Model Risk Management Principles for Banks	5 principles for effective MRM
Singapore	Report by Veritas Consortium	14 principles that promote Fairness, Ethics, Accountability and Transparency
U.S.	Algorithmic Accountability Act of 2022	Assessment of tools used for making critical decisions
Canada	Discussion paper, Industry letter	Principles of Soundness, Explainability, Accountability
	Bill C-27	AIDA: Mitigate risks of harm and biased output