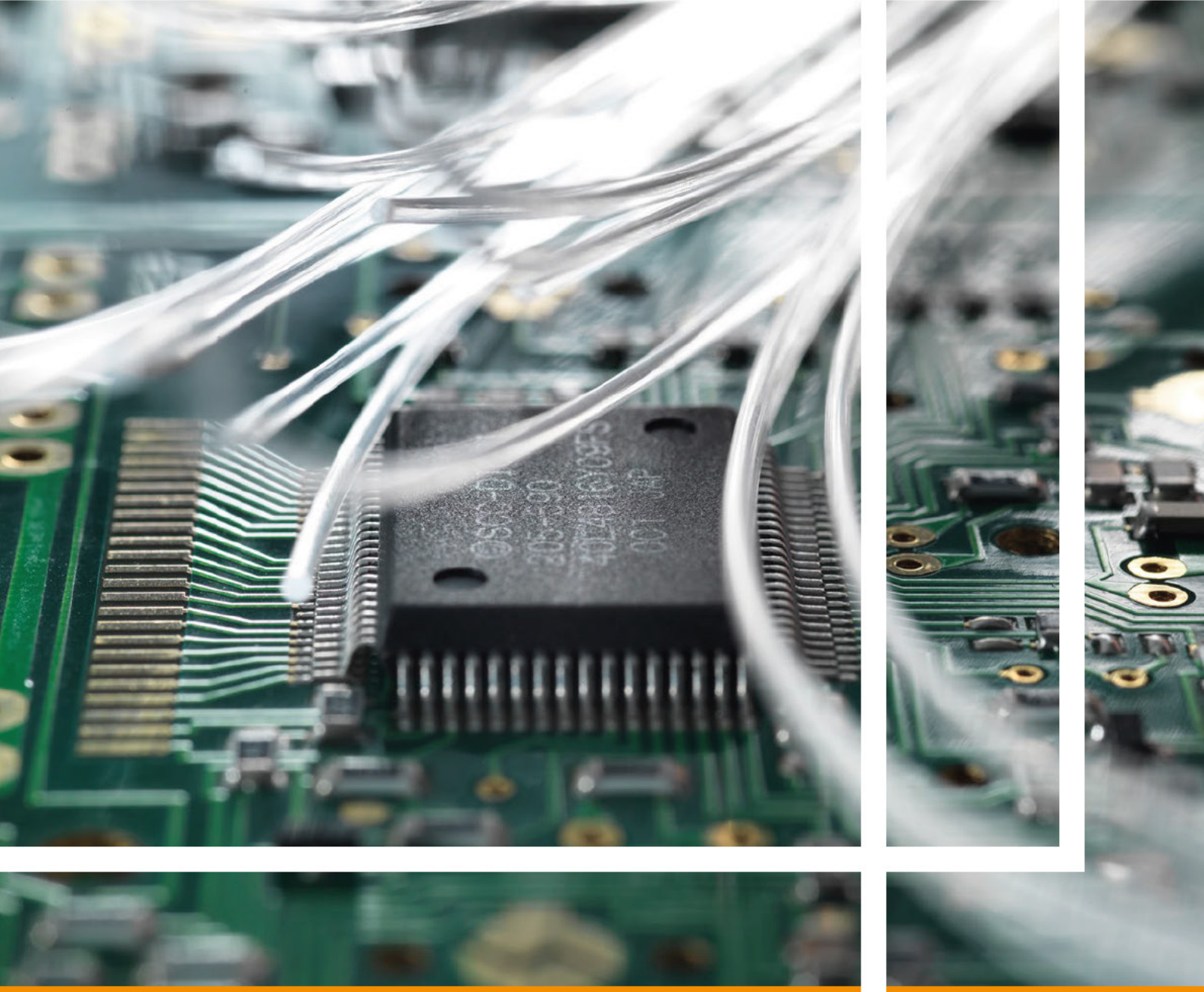


Model Risk Management of AI and Machine Learning Systems

pwc

Contents

Abstract

The purpose of this document is to present a model risk management approach for applied artificial intelligence systems. It reflects the nascent AI regulatory landscape and its expected near term development. AI systems are found to be consistent with broader definitions of 'models' applied in the financial industry, e.g. in regulatory guidance on model risk management, and therefore the latter are perceived as applicable, provided that extensions are made that address AI-specific risks. In this paper we review in detail the risks specific to AI, recommended approaches to deal with these risks, and the necessary adjustments to an existing model risk management framework.

1 Introduction

Consider the situation where you are buying insurance on a comparison website. After you receive your market quotes, you decide to go through the site again asking for more cover. To your surprise, you receive lower quotes! This is a real example, where the insurance companies' algorithms have been calibrated to perceive those requesting more cover as less likely to make a claim [64]. But this raises questions of fairness: is it fair to charge customers more for less cover? Also, as this 'feature' becomes well known, the consumer may change behaviour so that the previously observed pattern is no longer relevant. This is just one simple example of the difficulties that can arise from predictive models that 'learn' from all the dimensions in a historical data set.

The field of artificial intelligence ('AI') has seen significant progress in the last decade, with former frontiers of the technology having been exceeded exponentially; there are numerous examples, including handwritten digit recognition which was the frontier until the late 2000's vs. present day multiple object detection in live stream video [55] and basic bag-of-words language models vs. multi-task, multi-language deep learning NLP algorithms [56]. While previously limited primarily to the academic domain, nowadays the technology is being increasingly adopted in production settings to solve diverse business problems.

Machine learning ('ML') is a subdivision of the broader artificial intelligence field. It explores algorithms which can **learn** patterns from raw data and which run on computer hardware [57]. Arguably, the majority of contemporary **applied AI** use cases are relying on some sort of ML algorithm, resulting in the ML and AI terms often being used **interchangeably** in business contexts, with the term AI having become somewhat more common. In the present document we adhere to this prevalent convention to maintain alignment with multiple documents regarding regulation (e.g. [5,6,7,13,20, 27, 34, 35, 38, 43, 44]), remaining mindful of the differences in the strictly technical sense.

ML encompasses multiple algorithms, from traditional ones which have been in production for decades (e.g. linear and logistic regression, decision trees), to support vector machines, random forests, gradient boosted machines, various clustering methods, reinforcement learning algorithms and neural networks, with this list being far from exhaustive. Many of the recent successes in AI have come from a machine learning approach referred to as **deep learning**, which relies on neural networks with multiple hierarchically stacked processing layers (hence the term 'deep').

There has been evidence for increasing adoption of AI in production in multiple industries, with the financial services sector standing out as one of the early adopters [41]. With the growing maturity and broader application of the technology in business settings, there is an increase in the impact of AI-specific risks that naturally accompany its benefits. While capable of bringing massive improvements in efficiency and solving previously unaddressable tasks, AI does present a number of specific risks that need to be managed. Amongst these are increased complexity [41,43], possible privacy and data rights implications [6,7,19], potential to propagate or amplify existing biases [6,17,41,43] and difficulties in explaining the reasoning behind the recommendations of algorithms [6, 19,28,35,36,41,43]. AI adds a new area for model risk management, the practice of which is just starting to be developed.

In this document we review what these AI technology specific risks are, how they can be addressed and whether existing model risk management practices can be leveraged for the purpose. This is not done in isolation, but rather takes account of the opinions publicly expressed by relevant policy makers and policy advisors from the UK and the EU. On the one hand this allows us to draw from a broader set of views and research, and on the other to align with expected areas of scrutiny from regulatory bodies.

Whilst AI is a broad field encompassing mathematics, computer science, statistics, information theory and neuroscience, the scope of the present AI model risk management (MRM) document is narrower, focusing on *applied AI*.

We define this as models and applications used for practical decision making and control in production settings, functioning under the existing (and anticipated in the near term) regulatory environment for AI.

The review is to an extent oriented towards the financial services (FS) sector, based on its deep experience with quantitative modelling, as well as its status as an early adopter of AI. Nevertheless, good model risk management concepts are largely industry agnostic and most of the conclusions and best practices are applicable to a wide spectrum of other domains.

The paper is organised as follows: in the next section we review the impact of AI in financial services (FS) and outline why the latter has become a leading sector with respect to model risk management; next, we discuss the elements of effective MRM frameworks. Subsequently, we review the degree of applicability of existing MRM practices, as well as their limitations, with regards to their use for AI systems. Then we go on to discuss distinguishing features of AI systems, specific risks and possible ways to address them. Finally, we present a brief summary of our conclusions.

2 Impact of AI in financial services

The financial services industry, leveraging its experience in quantitative modelling and model-assisted decision-making, has been one of the early adopters of AI. A recent survey on ML in UK Financial Services conducted jointly by the Bank of England and the FCA [41] confirms that the technology has been increasingly adopted in the sector and that in many cases development has passed beyond the initial proof-of-concept or exploration phase. Two thirds of respondents are reported to already use the technology in some form and an almost equal proportion of applications are found to be in a 'live' stage of deployment [41]. Considering that AI technology builds on and often presents a more powerful extension of traditional statistical modeling techniques that have been used in the FS industry for decades, it is often used to complement (and sometimes substitute) existing 'traditional' models. It can also bring value in new areas where conventional approaches are not sufficient (e.g. voice assistants, multiple tasks in the natural language processing space and computer vision, and generative models [62,63]).

Historically, the FS industry has been operating under intensive regulatory focus with respect to model-based decision making. To address the risks from the wide use and potential material impact of models, especially after the 2008-2009 recession, the sector has been subject to strict and comprehensive regulatory scrutiny around 'Model Risk Management' (MRM). This has led to the industry gaining **pioneering** experience in MRM and developing regulator-driven mature MRM frameworks, the gold standard of which has been the Federal Reserve's 'Supervisory Guidance on Model Risk Management (SR Letter 11-7)' (SR 11-7) [3].

With the growing adoption of AI, however, organisations from the FS sector increasingly need to manage the risks which are concomitant with diffusion of the technology.

3 Model Risk Management as a mature discipline

Prior to the publication of SR 11-7 most banks carried out some form of controls, including independent validation, on those models that impacted the balance sheet (e.g. derivative valuation models), or regulatory capital (e.g. credit-risk or market-risk models). SR 11-7 widened the scope expected from banks to all models used in some way to inform business decision making, and introduced the concept of **model risk** as a risk to be managed like other well known risk types.

Since the publication of SR 11-7, other regulators have issued guidance with a consistent approach to MRM, and now most large banks and many smaller banks have designed and implemented MRM frameworks consistent with these regulations, generally including the following elements:

1. Developer testing and documentation (1st line): The onus is on the model developer to provide justification of methodology choice and evidence of testing before the model goes live. It is also expected that the model is well documented, including technical details and key model assumptions.
2. Independent review and validation (2nd line): All components of a model should be reviewed by an independent party. For material models, an independent function within the model risk team should carry out detailed independent review and testing. The inevitable conflicts of interest that arise in model development gives rise to the important principle of **effective challenge**, which is an important driver of the 2nd line review.
3. Both the 1st and 2nd line work should focus on the **conceptual soundness** of the model, highlighting model assumptions and model limitations, considering the pros and cons of alternative approaches, and carrying out thorough testing of the model's behaviour (e.g., sensitivity analysis, stress testing, limiting cases, and back-testing where relevant).
4. Periodic review and ongoing model performance monitoring, to confirm that the model continues to perform as intended, and that the most recent validation is still sufficient. 'Outcomes analysis', such as backtesting a statistical model, is particularly relevant here.
5. Model Inventory: The key information on each model is stored in a database, and accessed through an inventory tool. The keystone of a functioning MRM framework, a modern inventory tool will process the workflow of model approval, and aggregate model information into model risk reports.
6. Model Risk Tiering: This is necessary to allow for effective prioritisation, so that the majority of model risk resources are spent on models with the most material model risk.
7. Model Usage controls: Ensuring that the intended use of the model is aligned with the actual use.
8. Model governance and model risk reporting: This requires effective model committees where groups of models of the same type are approved, and model issues can be formally discussed between all stakeholders. This can then be the main conduit for reporting model risk metrics to senior management. It is crucial that senior management are aware of the key model risk issues and material sources of model risk in the firm.
9. Vendor models are treated to the same standard as internally developed models, with requirement for vendor to provide documentation, justify model design and show evidence of testing.
10. Internal audit role (3rd line). Regular inspection is required to ensure there is a consistent and effective approach to MRM across the firm.

These elements work together to improve the standard of modelling across a firm, reduce the number of material modelling errors, and ensure that models are used appropriately, with full awareness of the model limitations.

4 Applicability and limitations of existing model risk management frameworks

Based on a review of the definitions of 'AI', '[AI] algorithms' and 'models', outlined in multiple relevant sources including SR 11-7 (models), the Information Commissioner's Office [5] (AI), the EU High-Level Expert Group on AI [42] (AI), UK's Centre for Data Ethics and Innovation [17] (AI/ML), the Bank of England [2] (models), and other sources and relevant literature like [1] (AI), the following can be inferred:

- The perceived consensus definition of AI is fairly broad and seemingly encompasses: systems which include automated decision-making and recommendation; machine learning of any type (supervised, unsupervised and reinforcement), including deep learning models; symbolic AI systems; both software-based systems and physical robots; systems working in the areas of natural language processing; computer vision; audio perception; generative models as well as various agents for dynamic control. All of the above seem to be considered in scope by default. Note that this can include systems from the whole spectrum of machine learning, from very traditional linear models to highly complex deep learning algorithms, or reinforcement learning agents.
- The majority of the systems described above exhibit a common pattern: they consume data, process it and finally produce an outcome – a recommendation, a class, a predicted value, an action to be taken, or other information which can further inform decision-making or control tasks.

Importantly, this pattern is very consistent with prevailing definitions of a 'model' in the banking industry.

For example, while the Bank of England expects entities to devise their own definition of a 'model', they recommend all 'Calculation methods or systems that are based on statistical […] assumptions' and 'Calculation mechanisms used to transform a set of parameters or values into a quantitative measure' to be taken into consideration [2]. In addition, the Federal Reserve's SR 11-7 [3] considers 'quantitative method[s], system[s], or approach[es] that apply statistical, [...], or mathematical theories, techniques, and assumptions to process input data into quantitative estimates' to be models. SR 11-7 also argues that a model is considered to have an input component (where data is fed in), a processing component (where calculations and transformations are carried out) and a reporting component that provides outcomes or recommendations.

This definition is strongly consistent with how machine learning models function.

There are solid arguments for extending the validity and applicability of MRM best practice to AI systems. They stem from:

- The fact that MRM guidelines designed to abstract away the inner workings of the subject models and are technology agnostic; and
- The high degree of consistency of most AI systems with the definition of 'a model' according to the scope of FS regulatory frameworks [2,3].

However, models closer to the upper bound of the AI technology complexity spectrum, such as deep learning systems, have designs, inner workings and behaviours which are quite elaborate and often differ in important ways from 'conventional' modelling techniques. Further, AI systems are found to exhibit a distinguishable set of specific risks [6,7,19,28,35,36,41,43] that are not primary issues for most existing models.

Consequently, existing MRM frameworks cannot be mechanistically translated to the new breed of models without addressing these distinctive, technology specific risks which need to be controlled. **Therefore, the generic MRM 'best practice' frameworks are arguably a necessary, but not sufficient, starting point for AI model risk management**. In support of this notion, we understand there to be a broad consensus between views expressed by regulators and policy-makers [6,24, 58], policy-advisors [16,17,18,19,20,43], as well as industry [41] that AI can pose some unique challenges not encountered in traditional models (or at the very least that it can amplify a number of the existing risks seen in traditional models).

While traditional MRM frameworks would seem to provide a solid foundation for application to AI systems, to ensure full applicability they need to be revisited and potentially enhanced to address the distinctive risks stemming from such systems.

5 AI systems – specific risks and means to address them

As described above, gold-standard MRM principles are an appropriate way to manage the risks of AI systems, provided that adequate updates for AI-specific risks are made. Based on our understanding and the opinions expressed by relevant policy-makers and policy-advisors, we note the following main risks and issues that are specific to AI systems:

- One risk which is more pronounced in AI systems and thus tends to attract scrutiny is the dynamically evolving nature of the technology (new techniques continue to be generated at a rapid rate) as well as its relative complexity. This can make it harder for practitioners to have a comprehensive understanding of pros/cons of different techniques, creating implications for risk awareness and mitigation.
- Another primary issue is the area of model interpretability – the ability to explain the reasons behind the recommendations of a model; this is also closely related to model auditability and accountability which are other areas of high focus for policy-makers and policy-advisors.
- Further areas of importance are model bias & fairness, which while not exclusively AI-specific, can be propagated or amplified by AI systems.
- AI systems often **rely** heavily on huge amounts of data which can give rise to privacy, IP and data security risks. While models in other areas can be built on the basis of known principles or established mathematical relationships, AI models need a lot of data to **learn from**. This means that the use of this technology is closely related to handling large amounts of data which brings all of the data relevant risks – privacy, IP rights, suitability of the training data, etc.
- Businesses or divisions which might not have been historically engaged with data-intensive activities may need to build such capacities and risk management tools along the adoption of AI in their regular operations.
- Another area to consider is the frequent use of vendor or pre-trained models, which bring risks related to model design, transparency, and the relevance of training data. For such third-party tools, other risks such as bias and fairness may be harder to track and control.
- There are also AI-specific model risks to mention around 1. Stability (i.e. model that is generally accurate, non-biased and explainable, but gives inconsistent results in edge cases/ corner cases), and 2. Security (i.e. adversarial attack, model theft etc.)

Finally, there is also the important topic of **continuous learning** which is frequently raised in discussions on AI specific risks. Such learning is presumed to include ongoing model improvement based on fresh data from the AI system's continuous use and to take place **automatically** without human intervention (i.e. models to have in-built capacity to self-improve). While for some models (notably – neural networks) it is fairly straightforward to build a pipeline which continuously (or periodically) further trains or fine-tunes them, the problem of learning the new information without 'forgetting' the old is non trivial and hard to solve even at a theoretical level [60]. While workarounds do exist, for example 'experience replay' [61], for most use cases, true continuous learning models are still a problem of more scientific than applied practical significance, so we do not include them here as a present risk from the use of AI.

In the following subsections we discuss the more material of these risks, and how they can be addressed by using an enhanced MRM framework.

Model complexity and implications for risk awareness

Certain AI models, especially those deployed to fulfil high-value use cases, can be highly complex. In addition, the AI area is evolving very fast and brand new models, concepts and architectures emerge on a frequent basis, making it non-trivial to keep up with technology's progress. Furthermore, some commercial products offer 'automated' ways to train models which may allow modelling with very little, if any, understanding of risks, advantages and disadvantages of the algorithms deployed. Also, there are an increasing number of publicly available pre-trained models in the NLP and computer vision areas, which can be freely used or repurposed. These phenomena bring challenges in terms of organisational awareness and understanding of model risks.

One of the main considerations in model evaluation is conceptual soundness and fitness for purpose. Unlike many traditional models that represent well-established assumptions of actual cause-effect relationships (e.g. pricing models, physical processes models), AI systems **learn** assumptions about those relationships from exposure to data which renders their inner logic not readily observable. The high complexity, fast evolving nature and ability to use vendor or 'out-of-the-box' models often makes the evaluation of conceptual soundness harder than is the case for traditional models. While the details will vary based on case specifics, there are some overarching principles of conceptual soundness which are universally valid and apply to AI/ML systems well.

Firstly, using 'automated' machine learning in production without sufficient understanding of the nature of the models being trained, is not to be recommended. Maintaining internal capacity to understand the nature of the risks of models in use is essential.

Secondly, there are domains with elevated risk and degree of scrutiny, for example facial recognition, using data from social media or models used in the human resources areas [65]. Careful consideration is warranted on when to use AI technologies in those domains as the outputs can be strongly context dependent [66]. The fact that the technology is theoretically able to perform a task does not mean it necessarily needs to be applied for that task. Whether to apply AI for a specific problem is a broader business question which encapsulates commercial, regulatory and customer acceptance considerations which go beyond the technical capacity of the technology to perform a task.

It is also noteworthy that present day AI is far even from basic levels of intelligence in the biological world and machine learning algorithms can not be expected to solve many abstract tasks. Conversely, AI has an excellent, often superhuman, capacity to solve problems in pattern recognition, classification, perception (e.g. vision tasks or audio processing), make sense of high-dimensional data as well as to perform tasks which constitute relatively narrow slices of what we would understand to be human intelligence. Therefore, AI models can offer great solutions for multiple constituent components of a variety of decision making processes. But they shouldn't be expected (or advertised) to exhibit capabilities they are presently unable to possess. There is a natural human tendency to attribute too much confidence to an AI model, or any computational tool, because of the speed and accuracy of its operation, even leading us to attribute the property of 'understanding' to the tool.

To conclude, organisations need to make sure they are sufficiently prepared to understand and handle the risks of the technology. A practical approach for evaluation of an organisation's preparedness includes answering the following questions:

- Are they able to explain how the selected model or technology works and what its relative advantages and disadvantages are?
- Are both model owners and senior management aware of the risks and limitations of the technology?
- What is their level of experience with this technology?
- Have they seen solid industry and academic evidence for application of the selected technology for the case at hand?
- What part of the model creation do they control (e.g. building it from scratch, repurposing existing models from similar cases, using vendor 'accelerator' products, using 'out-of-the-box' pretrained models or vendor models)
- Do they have control over the data the model has been trained on and are they confident the training data is representative of the target population?
- Have they evaluated the relative cost of added complexity and decreased interpretability vs. the potential gain in performance, mindful of the context of model use?
- Is management content with the outcome of this evaluation in a broader business sense (not purely the degree of model technical performance)?

To address the questions above, even in the case where a mature MRM framework is in place, firms would need to establish internal capacity to fully understand and evaluate AI specific risks. In practice, this could mean categorisation of those models as belonging to the AI/ML subset, adding them to the model inventory and establishing dedicated teams comprised of AI/ML subject matter experts to handle those models, across all three lines of defence.

As this technology is becoming increasingly prevalent, it is more frequently used in domains and industries that have not had historical exposure to model-assisted decision making. In such cases, a formal model risk management approach might not exist at all. Consequently, organisations who lack a sufficiently comprehensive MRM policy but are now in a situation to use AI/ML for material business decisions or processes would need to develop such policies **de novo**. Drawing from the experience of industries with leading experience, e.g. the FS sector, can accelerate and greatly assist these efforts.

Thirdly, in terms of governance and senior management oversight, the leadership of organisations who use AI systems in production would need to establish effective oversight, risk management bodies (e.g. committees) and build full awareness about the risks and limitations of this technology.

Transparency (interpretability)

The ability to explain the reasons behind machine learning models' decisions has been a central theme in multiple policy-maker and industry sources [5,6,28,34,35,36,41,43,58]. Interpretability naturally comes as an AI/ML specific issue due to the very nature of this technology. Most traditional models rely either on existing and fully known, hard-coded mathematical relationships, or, on readily explainable **learnt** but **linear** statistical relationships. Conversely, AI systems in most cases provide **learnt, non-linear** statistical relationships which are usually encoded in a very high number of parameters, with the number often in the millions. This makes the internal functioning of such systems naturally hard to explain in traditional, low-dimensional terms.

Interpretability is identified as essential because of its relevance for model accountability, auditability, control of bias and ensuring fairness. There seems to be a broad consensus that:

- Model interpretability is not free and almost always comes at a price which is reduced performance
- Trade-offs need to be made when defining the desired level of explainability
- The appropriate level of interpretability is strongly dependent on the use case and there cannot be a one-size-fits-all approach
- Full explainability may not be possible in certain situations
- There is no established 'best-practice' for delivering model explainability

Here we outline a practical perspective for the problem, mindful of the views expressed by major policy-relevant institutions, and respectful of limitations and specifics of the techniques available.

We focus primarily on nonlinear algorithms (or 'black-box' models) rather than linear ones for several reasons. Firstly, linear models like decision trees and logistic regression are explainable by design. Secondly, they may be expected quite often to belong to lower model risk tiers. Thirdly, the explanations for such models are straightforward, consistent across cases and do not require arbitrary tuning of various hyperparameters which in other cases can influence the results greatly. While we note that some traditional statistical models (e.g., linear regression) would come under the broad definitions of 'AI' in regulatory policy documents [1,5,42], we do not find the term AI to be particularly applicable or helpful for these cases. Further, for those models explainability is not an issue, and existing MRM frameworks already cover them without the need for extensions.

On the other hand, nonlinear ML algorithms like random forests, gradient boosting machines and all kinds of neural networks, as well as the majority of practicable reinforcement learning agents, do not naturally lend themselves to human-accessible explanations. This means that most problems in high-dimensional regression and classification settings, natural language processing, computer vision or dynamic control domains are quite opaque, unless solved with linear algorithms (which often is not possible at all or comes with severe limitations in performance). To accommodate those models, a number of model agnostic 'black-box' explanation techniques have been developed and are being increasingly used in practice. They frequently rely on variations of a common approach for which a high-level description is provided below.

Many such techniques broadly consist of getting a model which is already trained and then treating it like a black box, experimenting with its behaviour under controlled variation of its inputs. Thus an approximation is derived for how the model would react to a given set of inputs or given some changes to the provided set of inputs. Some of the more popular representatives of this broad class of techniques include partial dependence plots [47], individual conditional expectation (ICE) [47], permutation feature importance [47], local surrogate models (LIME) [48], deriving rules (Anchors) [49], Quantitative Input Influence [50], Shapley values (SHAP) [51].

There are alternatives like training global surrogate models [47], which are linear and are trained to do the same task as the non-linear ones; explanations for the linear ones are then used to approximately interpret the hypothetical reasoning of the nonlinear model.

There are even more complex approaches which are aimed specifically at deep learning models and work at a lower level. Examples of these are examining neural activations [47], using an attention mechanism to explain exactly what the model is looking at [52], or training a model with an 'explainer' layer jointly so that predictions are intrinsically explained [53].

All of the techniques described so far can provide information about the behaviour of a black-box model, although the level of their reliability is not guaranteed and is often variable across cases. While a detailed technical review of pros and cons of available explainability techniques is beyond the scope of this document, a few points are worth mentioning.

Explainability techniques which rely on tweaking the inputs to measure how this affects the output present some challenges. First, many of them disregard the interconnectedness between different features and assume they are independent. Also, they may often rely on unrealistic examples which emerge when features are mechanistically varied. They also frequently require some arbitrary setting of parameters which can greatly influence the outcome; usually, those parameters are chosen based on rules of thumb and may require multiple runs.

Meanwhile, training a global linear surrogate is stable and easier to use by a non-expert audience in the production setting. However, it relies on the assumption that the nonlinear model is approximately consistent with the decision-making logic of the linear one which is often not the case. Moreover, many cases in natural language processing, computer vision or other complex domains are not solvable (well) with linear models at all.

Finally, the very advanced approaches like using neural networks' attention layers [52] or training explainers jointly with the target models [53] may be popular in the scientific community but they usually require a very advanced level of expertise and are often hard to implement. Therefore, they can only be used in select cases where the context warrants it.

Mindful of the above, in the following paragraphs we share some practical considerations for addressing the model interpretability issues.

- First, it is sensible to apply the model risk tiering concept when evaluating the desired level of explainability. High-risk models are worthy of deeper consideration: models with an elevated level of risk and/or models which are related to decisions about individuals (especially decisions which can be impactful), to privacy, to potential predetermination of individual's choice; models using sensitive variables (even if using them is formally legitimate); models related to decisions in healthcare, human resources, allocation of benefits. All of these are examples that should drive a higher model risk rating, and they should meet more stringent explainability expectations. Existing MRM policies need to be updated to ensure they take into consideration the relevant risks when rating AI/ML models and deciding what the desired level of interpretability emphasis to be in any given case.
- It is also important to consider what the potential audiences of the explanations would be (e.g. regulators, individuals that are subject to model decisions, experts or non-experts, auditors, organisation's senior management); the format and depth of the explanations need to be proportionate to the risks involved and appropriate for the expected audiences. Serving diverse audiences might require more than one type of explanation (e.g. an in-depth and a high-level one) if the spectrum of users of the model explanations is wide.
- Using AI on a significant scale will require building internal capacity to both produce and interpret model explanations. This may also include building an internal methodology for explanation of different types of models and embedding it as a part of MRM policies, so a standardised approach customised to the specific models at hand is available. The latter would allow internal best practice to spill across departments and functions, as well as comparability of model review outcomes.
- The outcomes of model explanation techniques are not to be taken at face value and need to be produced and critically reviewed by staff familiar with the strengths and weaknesses of the techniques. This is one of the reasons for resourcing the three lines of defense with staff that have sufficiently deep and relevant AI/ML expertise.
- Senior level audiences might need to be briefed with summaries of the outcomes and major conclusions from model explanation and validation work rather than be expected to review deep technical details directly.
- Application of more than one model interpretability technique is preferable in the vast majority of cases. For each model (or model type) a set of techniques (e.g. 2 or 3, or occasionally more) can be considered, preferably with techniques whose pitfalls do not overlap. As a baseline, using a global surrogate linear model can often be considered, if the case allows. A baseline global view on generally important variables and inputs can be provided this way, and also major inconsistencies or unexpected behaviours can be captured. Furthermore, this is an explanation technique which is consistent across cases and can be later used as a benchmark vs. more complex techniques. It also provides a very good view how much (if at all) a target black-box model is superior to the linear one. If the difference is not sufficiently large, the use of a black-box model might not be warranted in the first place.
- At least one additional technique, e.g. LIME [48], SHAP [51], Anchors [49] or other can be used, comparing the outcomes with the baseline results from the linear surrogate model. Particular attention may be warranted in cases where the linear and nonlinear models disagree; ones which are borderline between categories, as well as whenever there is a difference in the sign, or substantial difference in magnitude of variable impact in the two model interpretation techniques.
- An evaluation of the relative cost of inaccurate model decisions would inform the accuracy vs. interpretability trade-off. The reasoning behind the final choice of model, in terms of the extent of interpretability, should be appropriately documented.

- Once model explanations are derived they can be sense-checked with subject matter experts from the relevant domain. Explanations not making sense to SMEs are important red flags.

- Finally, there are models which are not naturally good targets for existing interpretability techniques. Models in computer vision, reinforcement learning and content generation are often either not explainable or the derived explanations don't make much sense to non-technical audiences. In such cases, organisations may need to consider whether using AI in the particular context is warranted with respect to external requirements and their own model risk appetite. If using such models is found warranted (e.g. scanning documents is a computer vision application which doesn't meet much external pressure for explainability), using some of the more advanced explainability techniques like [52,53] during the development and validation phases may make sense, as they can expose poor model behaviour such as overfitting. Techniques like LIME [48] and Anchors [49] are technically applicable to NLP and computer vision models, but the output needs to be interpreted carefully.

- A special case is the application of third party (vendor) models. If an organisation is accountable or liable for the performance of a model, even if it is created by a third party, virtually the same model interpretability outcomes can be required for those vendor models too. If model explanations are provided by the vendor, they need to be critically evaluated and accepted only if considered sufficiently comprehensive and reliable as vendors would inevitably have an inherent conflict of interest (i.e. whether they can be independently critical of a product they have both created and are trying to sell at the same time). Deriving bottom up explanations for vendor 'black-box' models is possible, though it normally requires full access to the model which is often not available. In general, AI/ML vendor models need to be treated in the same way as internally developed ones in an MRM sense, and need to be added to the model inventory and go through the relevant governance process.

Finally, another potential issue related to model explainability is preventing uncontrolled use of the model by external parties. If external audiences have the ability to experiment with a model, its sensitivity to various inputs can be derived, allowing for potential 'gaming' of the system (e.g. multiple applications for the same product until figuring out which features the model considers advantageous) or finding ways for adversarial attacks (e.g. in biometric authentication cases). In general, adherence to a 'need-to-use' policy for each inference instance is a sound approach. In extremis, this technique can also be used to 'steal' models by recreating model parameters through recurrent querying.

Model bias and fairness

Model bias & fairness is arguably the other topic that appears most prominently and consistently as an area of focus for policy makers and regulators [6,17,41,43,58].

While definitions vary, generally a model is considered biased if it is generating decisions that treat distinct groups of entities differently without an objective reason, or it is making decisions based on inputs which cannot be reasonably expected to matter in that particular case.

The opinions of major policy makers, policy advisors, and regulators appear to agree that:

- There is no universal definition of fairness;
- Bias in decision-making is not a new phenomenon specific to AI and ML;
- AI and ML models can propagate and even amplify biases which have long existed in legacy processes and datasets;
- Fairness has multiple dimensions which are often in competition with each other (e.g. fairness to the individual vs fairness to the group);
- Optimising with respect to all of those dimensions is not usually possible and therefore informed trade-off decisions need to be made.

Below we discuss a practical approach for addressing the bias and fairness aspect of AI model risk.

First, it needs to be decided what is in scope. It is evident that the concept of fairness is applicable mostly in cases where a model's decisions are likely to impact **individuals** or **firms** and especially where there is a perceived asymmetry of power between the organisation and the entities affected by the AI system's decisions. For models like the ones used for optical character recognition, translation, document information retrieval, asset pricing, computer games, and many others, the bias and fairness topic is not generally relevant unless their outputs are likely to, directly or indirectly, affect individuals significantly.

Whether or not a model needs to be subjected to detailed bias and fairness review can be decided during the model **risk rating** evaluation process, making the 'bias & fairness' consideration another area of risk which MRM processes need to address. Questions relevant for this process may include:

- Does the model provide decisions which can have a non-negligible effect on the life of individuals? Examples are decisions for consumer loans and mortgages, life insurance access and pricing, access to treatment and medications, paroles, automated video surveillance, scraping of social media profiles, employment decisions, etc.
- Does the model limit individuals' free choice?
- Does the model limit the human rights of individuals?
- Is a model applied in an area for which significant asymmetry of power can be considered?
- Is the model operating in an area where historically there have been issues with accessibility, exclusion and fairness?
- Does the model use any 'protected variables' (these are variables which are prohibited or regulated by law, or more broadly of sensitive character: religion, gender, disability status, healthcare records, etc.)?
- Does the management consider there might be other aspects which make the bias & fairness concept relevant for the model at hand?
- Has the broader impact of the model been considered (e.g. if it feeds data to downstream models) and does the system as a whole meet the criteria above?

If a model is found to be out of scope, further practical considerations may include: Highlighting the reasons for such a decision in the model's documentation; Revisiting the 'Is it in scope?' question during regular model risk reviews.

If a model is found to be in scope conducting the following additional steps should be considered:

1. Taking actions for protected or sensitive variables;
2. Addressing model bias with respect to input variables.

The first of these steps includes:

- If any applicable law or regulation prohibits the use of an attribute in a particular context, this variable should be taken out in model application (inference) settings. While this is universally valid regardless of the classification of a model, if formally admissible it might make sense to keep track of such variables to **allow** measuring and preventing unfair bias or discrimination.
- Next, a check whether some formally allowed but generally sensitive variables are used (e.g. age, gender, religion, etc.); some variables may be considered sensitive by some entities and not sensitive by others, or sensitive in some contexts and not sensitive in others. In cases where there are no formal restrictions, whether or not the use of a particular sensitive variable is acceptable for the organisation is primarily a matter for management judgement
- The benefit from using sensitive variables needs to be weighed against the reputational and other intangible overhead they bring. In some domains, such as healthcare, variables like age and past medical history might be necessary for an informed decision; in others, using them might be contentious or prohibited.

The second step, addressing bias with respect to input variables, is a complex process which requires a lot of judgement and is closely associated with the topic of model explainability, as interpreting a model's decision is usually a key step in detecting bias and measuring fairness.

An example workflow may include the actions described below.

First, getting global and local model interpretations and reviewing the impact of each feature (or at least the most impactful ones if their number is very high) might be considered, in combination with checking whether the sign or magnitude of their impact makes sense with respect to expected behaviour. A next step might include deciding with respect to which attributes bias & fairness need to be measured (e.g. age, gender). While some of these sensitive variables might not be used in the model at all, if they are available (e.g. from diversity surveys) and if it is formally and ethically permissible, checking whether decision-making systems are fair with respect to those attributes might make sense. A subsequent step might include evaluation whether the selected sensitive features are predictive for the particular decision making system as well as how impactful they are (e.g. marginally, moderately, strongly). If they are impactful, it may need to be checked whether there is a fundamental reason for this to be so (e.g. age is often justifiably important for decisions in healthcare). If there is an undesired or unexpected disparity between any two groups of interest (e.g. male vs. female applicants), it may need to be considered whether they reflect historical biases in legacy processes and the training set, whether there are class imbalances, presence of some groups that are underrepresented in the training data sets.

Multiple approaches can be used for the analysis described above, including the predictive power of the respective variable, statistical significance, along with methodologies for measuring and removing disparate impact [54], for ensuring 'equality of opportunity' [59] or methods outlined in PwC Responsible AI framework.

In some cases, an identified disparity may not necessarily be attributable to deliberate unfair bias but the idiosyncrasies of the limited training set or suboptimal selection of the training data.

Removal of bias with respect to a particular variable may include not using it in the first place (not necessarily enough) and looking for proxy variables (e.g. whether the particular sensitive variable can be inferred based on a combination of the remaining attributes). For example, if gender can be very accurately predicted based on other features, just dropping it from the dataset won't solve the problem.

Techniques like the ones described in [54] and [59] allow breaking those proxy relationships or applying compensatory adjustments to the models, with some tunable sacrifice in model performance. However, mechanistic removal of detected disparity in outcomes between arbitrarily chosen groups can cause tensions between competing nuances of fairness (e.g. raising the price of services to everyone or decreasing acceptance rates vs. making the model blind for a particular feature). Moreover, removal of so defined bias by using statistical techniques which break relationships between inputs usually has another price component: decreased model accuracy. This may lead to more conservative acceptance criteria which then can lead to exclusion of some entities from access to the services they would have otherwise been granted. Therefore, mechanistic removal of the disparity of outcomes between arbitrarily chosen groups can cause collateral harm and removal comes at a price of reduced model performance which on its own can have fairness and inclusivity implications. Rectification of one dimension of fairness often may lead to deteriorating others. The resulting trade-offs are management decisions which reflect organisations' ethics principles and go beyond technical terms and parameters.

Finally, the agreed degree of removal of identified biases and the rationale behind trade-off decisions should be thoroughly described in model documentation.

To ensure appropriate handling of the bias and fairness issue and senior management oversight, MRM policies need to recognise it as a distinct problem and include well-defined processes and distribution of responsibilities for addressing it.

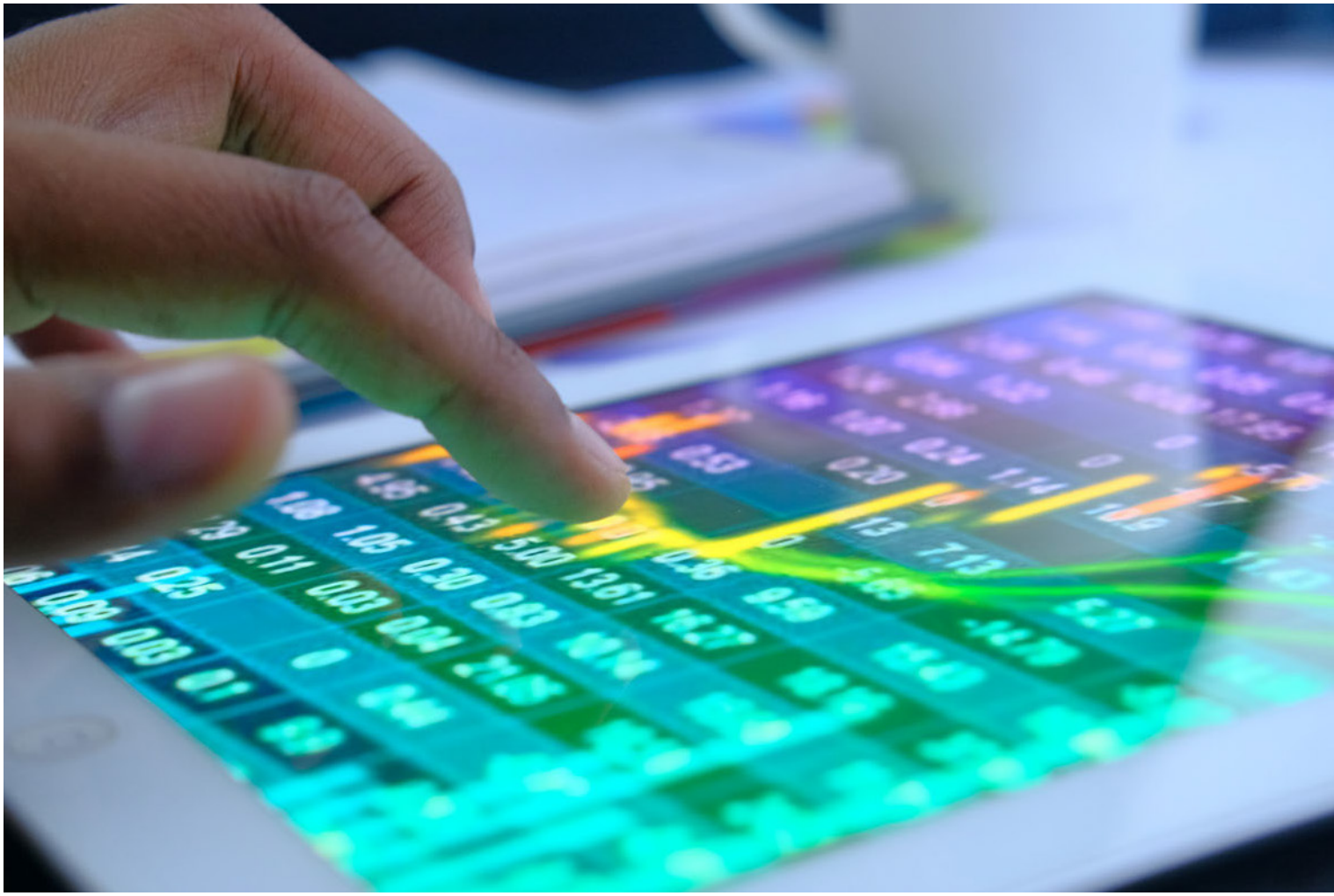
Third party models and transfer learning

Using third party vendor models has been increasingly prevalent in the AI space. Another issue is the use of pre-trained generic models, which become quite important due to the huge amount of training data, training time and computational resources that large computer vision or natural language processing systems require.

The standard MRM approach to models designed and trained by third parties is that they need to meet the same requirements and answer the same questions as if they were developed internally. This also includes open-source pretrained models in the NLP and computer vision areas which leverage access to the enormous amounts of data that some organisations (e.g., large technology companies) have.

Very often, the possibilities to scrutinise such models are fewer than those for internal models for various reasons, including: vendors not willing to expose their IP; vendors not motivated as they don't bear the ultimate risks; insufficient documentation provided by the vendor.

Extra attention might be necessary regarding bias (if applicable for the use case) and representativeness of the training data for the population to which the models are going to be used in production. In such cases the training data inevitably comes from outside the user organisation, and so this will need evaluation and challenge.



Data

AI systems are heavily dependent on significant amounts of data, making developer and user organisations exposed to data-related risks, namely – quality, IP rights, personal data protection, security.

In addition, with the democratisation of AI and advent of multiple 'automated' machine learning platforms, it becomes increasingly popular to train and use such systems in an out-of-the-box manner. This allows AI to become the gateway through which organisations can (often inadvertently) expose themselves to new data risks, as well as compulsory regulation.

Generic data risks constitute a topic which is broad enough to warrant its own governance policy; this domain is also comprehensively covered by existing, non AI specific regulation like the GDPR.

Regarding AI model risk management, organisations need to acknowledge and consider those risks in their model governance frameworks and add them as a separate dimension of risk in the model risk tier evaluation process. MRM frameworks need to be updated, or to provide meaningful links, to generic data governance policies so that the 'data' dimension of risks of AI systems is appropriately addressed and brought in compliance with existing data regulation.

Nascent regulation

There is a lack of established overarching AI regulation and this is still subject to active debate and development, putting a layer of uncertainty to user organisations. We are presently not aware of any overarching compulsory regulation that is focused specifically on artificial intelligence in either the UK or the EU. However, such rules appear to be in the making [15,16,17,44,46, 65, 66]. Think tanks with the primary objective to advise governments on future AI policy-making have been established [12,15,23,45] and have already issued a number of draft or discussion documents reviewing the main issues and areas of interest (**these documents are listed in the Bibliography section below**).

To date, there appears to be a significant degree of convergence of the topics these organisations are focused on and their approach to the overall landscape. In general, they demonstrate a positive attitude towards the technology and recognise its potential to bring huge opportunities.

Some areas of difference between the UK and the EU approach are detectable albeit they might not necessarily translate proportionately into the regulation that emerges in the coming years. Our reading is that the UK institutions' approach, including that of organisations such as the ICO, the Bank of England and the FCA, tend to put more emphasis on developing recommendations, 'best practice' and guidelines, leaving existing sectoral regulation to extend to AI models without overtly planning for new sets of compulsory norms in the near term. It could be speculated that online targeting may prove an exception to this trend but as of the date of this document's publication, the CDEI's policy recommendations are yet to be released [16] and therefore the question remains open.

By contrast, the EU appears to be more focused on going 'beyond the voluntary guidance' [44] and establishing formal regulatory, oversight and enforcement norms for AI, including 'enhancement of institutional capacity' [44] where necessary. There is a declared preference towards a common approach at a Union level [44, 65]. Considering the issue is in the EU's top government priority agenda [46] it can be speculated formal policy-making actions are likely to take place soon; 'prior conformity assessments' with **future** formal requirements for 'high risk' AI systems as well as the possibility for formal procedures for testing, inspection and certification of such are also being under consideration [65].

It seems that both UK and EU institutions which have influence on AI policy-making favour outcome-based approaches on rule-making with respect to AI. It could therefore be inferred that very prescriptive or prohibitive regulation is currently less likely.

Despite the lack of specific compulsory norms on AI, there is already a relatively well distilled view on which aspects of the application of this technology governments and regulators are likely to be most interested in. This also extends to sectoral regulators like the ICO, the FCA and the PRA whose domains will have common interfaces with AI or are expected to be influenced by the technology. Their views can be indicative of how they intend to apply the existing sectoral regulation when AI is involved in the respective area.

Based on the observations above, we suggest it is important to develop, deploy and use AI technology in a way that is mindful of the current thinking and trends amongst policy-makers. MRM policies need to reflect the areas of expected regulatory scrutiny to ensure compliance in the coming years. This should help minimise the chances of making investments in technology and infrastructure which do not account for the most significant risks or likely regulatory developments. As policy-makers and regulators' views are in an evolving phase, organisations will benefit from staying tuned to any developments in the institutional guidelines.

Conclusion

We have presented a discussion on AI model risk management and outlined a practical approach to addressing the issues. Our suggested approach combines established model risk management best practice with additional technology specific considerations, and aligns with our understanding of prevailing regulatory thinking.

First we reviewed what existing MRM best practice there is and what parts of it can be leveraged for AI systems, finding that existing generic principles lend themselves very well to AI, provided that several enhancements and considerations for technology specific risks are applied. We also identified those AI specific areas and proposed actionable measures to address them.

We have argued in this paper that MRM is broadly applicable to most emerging use cases for AI. For those firms that don't have an MRM framework in place already, but have an expanding portfolio of AI applications, this should be the first step for controlling the risks of AI models.

We now summarise our recommendations on how the MRM frameworks commonly in place in large banks should be extended to AI models. We group this extension into the following 7 steps:

1. **Identification** of AI tools – All instances of AI applications should be placed in the model inventory and tagged as 'AI'. Some existing models may also warrant this tag. As there is no clear-cut definition of AI, the labelling should use risk-based criteria (i.e., does the model present the AI-related risks discussed in this paper) and expert judgement. The use of this 'AI' tagging of models is to allow for the following six steps to be applied to all AI models.

2. The **Model-risk rating** criteria should be updated for AI models to include specific AI risk factors, e.g., whether explainability is a problem, or if there are issues of bias/fairness, such as a model impacting individuals or using ethically sensitive variables.

3. **Bias/fairness** for those AI models intended for use in sensitive domains. It must be recorded how issues of bias and fairness are addressed, and this must be subject to independent challenge (typically from a 2nd line function). Standardised techniques must be defined and used to detect model bias with respect to sensitive factors, and to address these biases.

4. **Explainability** tools require sufficient oversight and governance. Where the Black-box nature of a model, or its lack of domain-specific assumptions, leads to a lack of clear understanding as to the reason for the decisions of an AI model, a defined approach is required to determine if interpretability techniques used are appropriate. There should be an SME review for each application of such techniques to test the suitability of approach. We recommend that firms build approved internal methodologies for explainability for each class of AI model, and embed these within the MRM framework; in some cases the techniques used for explainability should themselves be classed as models, with the relevant approval and controls.

5. **Suitability** of using AI for a given application. For each distinct use of an AI tool, a decision process should be recorded on whether this is an appropriate choice: Does the cost of lower interpretability justify the gain in performance over simpler linear models? Is sufficient expertise in place by users and risk management? Is this an industry-standard application? Are there data privacy issues? Etc.

6. **Ongoing monitoring** as an AI model may change behaviour as it learns new patterns from additional data. While ongoing monitoring of model performance is a well-established part of existing MRM frameworks, the fact that AI models are making decisions based on high-dimensional features that may not be obvious to the human eye means that the model results may qualitatively change, as new training data arrives. For this reason, the testing of model accuracy, as well as bias and explainability, need to be regularly tracked on a more frequent basis than might be the case for traditional models.

7. An **AI risk committee** should be established as an overlay on existing model governance. This would consider the overall population of AI models, and the implementation of the extended AI MRM framework. The committee should monitor the emerging risks of AI, highlight areas of weak governance and ensure that the use of AI is within the risk appetite of the firm.

In terms of resources, to ensure effective AI MRM, organisations using such systems need to maintain staff with relevant expertise across all three lines of defense and across the hierarchy vertical, including competent decision-making bodies.

Organisations with significant experience in general MRM and the respective level of model risk awareness should be able to adapt to AI system risks with fairly modest adjustments. Others, for whom this technology is a means to apply large-scale model-assisted decision-making for the first time and who are still to develop model risk-aware thinking, will have to embark on a longer journey. For the latter, drawing from existing best practice MRM could be especially helpful.

We can conclude that a holistic and high-quality system of AI model risk management should cover the entire lifecycle of the AI models, should include comprehensive independent validation and should include full organisational awareness of inherent risks and frontiers, as well as a formalised governance approach which explicitly covers the identified areas of AI/ML system specific risks.

While we acknowledge that this is formally compulsory only for entities from the financial sector in particular geographies, the outlined principles can be considered largely country and industry agnostic. Therefore, voluntary application in FS entities from other territories, as well as firms from other sectors, is likely to be of great utility for control of risks.

Finally, it was identified that many of the risk-relevant questions are not technical but rather managerial in nature and therefore require a significant degree of commercial judgement. Therefore, while automation plays an important role for model performance monitoring, the key components for robust AI system validation and risk management are still formalised governance policies and human resources with the appropriate expertise to implement those policies.

Bibliography

1. Law Library: Library of Congress. (2019). Regulation of Artificial Intelligence in Selected Jurisdictions. The Law Library of Congress, Global Legal Research Directorate. https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf. Retrieved on 31.10.2019 at 11:58

2. Bank of England. (2018). Model Risk Management Principles for Stress Testing. Supervisory Statement | SS3/18. https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2018/ss318.pdf. Retrieved on 31.10.2019 at 13:41

3. Board of Governors of the Federal Reserve System Office of the Controller of the Currency. (2011). Supervisory Guidance on Model Risk Management. SR Letter 11-7. https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf. Retrieved on 31.10.2019 at 14:01

4. Information Commissioner's Office. (2017). Technology Strategy 2019-2021. ICO. https://ico.org.uk/media/about-the-ico/documents/2258299/ico-technology-strategy-2018-2021.pdf. Retrieved on 30.09.2019 at 15:14

5. Information Commissioner's Office. (2019). Project ExplAIin. Interim Report. https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf. Retrieved on 30.09.2019 at 15:17

6. Information Commissioner's Office. (2019). AI Auditing Framework – An Overview of the Auditing Framework for Artificial Intelligence and its Core Components. https://ai-auditingframework.blogspot.com/2019/03/an-overview-of-auditing-framework-for_26.html. Retrieved on 30.09.2019 at 15:34

7. Information Commissioner's Office (2017). 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf. Retrieved on 30.09.2019 at 16.06

8. HM Government. (2018). Industrial Strategy: Artificial Intelligence Sector Deal. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/702810/180425_BEIS_AI_Sector_Deal__4.pdf. Retrieved on 01.10.2019 at 11.27

9. https://ico.org.uk/about-the-ico/. Retrieved on 30.09.2019 at 15:03

10. https://en.wikipedia.org/wiki/Information_Commissioner's_Office. Retrieved on 30.09.2019 at 15:04

11. https://www.wired-gov.net/wg/news.nsf/articles/A+call+for+participation+Building+the+ICOs+auditing+framework+for+Artificial+Intelligence+19032019161000?open. Retrieved on 30.09.2019 at 15:42

12. CDEI. (2019). Introduction to the Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/787205/CDEI_Introduction-booklet.pdf. Retrieved on 03.10.2019 at 12:51

13. HM Government. (2018). Industrial Strategy: Artificial Intelligence Sector Deal. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/702810/180425_BEIS_AI_Sector_Deal__4.pdf. Retrieved on 01.10.2019 at 11.27

14. HM Government. (2019). Centre for Data Ethics [and Innovation] (CDEI) 2 Year Strategy. https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2-year-strategy/centre-for-data-ethics-cdei-2-year-strategy. Retrieved on 03.10.2019 at 13:55

15. HM Government. (2019). Centre for Data Ethics and Innovation (CDEI) 2019/2020 Work Programme. https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme. Retrieved on 03.10.2019 at 14:31

16. CDEI. (2019). Interim Report: Review Into Online Targeting. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819169/Interim_report_-_review_into_online_targeting.pdf. Retrieved on 03.10.2019 at 14:33

17. CDEI. (2019). Interim Report: Review Into Bias in Algorithmic Decision-making. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819168/Interim_report_-_review_into_algorithmic_bias.pdf. Retrieved on 03.10.2019 at 14:36

18. CDEI. (2019). Deepfakes ad Audio-visual Disinformation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf. Retrieved in 03.10.2019 at 14:37

19. CDEI. (2019). Smart Speakers and Voice Assistants. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831180/Snapshot_Paper_-_Smart_Speakers_and_Voice_Assistants.pdf. Retrieved on 03.10.2019 at 14:38

20. CDEI. (2019). AI and Personal Insurance. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833203/AI_and_Insurance_-_WEB.pdf. Retrieved on 03.10.2019 at 14:40

21. Rovatsos, M.; Mittelstadt, B.; Koene, A. (2019). Landscape Summary: Bias in Algorithmic Decision-Making. CDEI. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819055/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf. Retrieved on 09.10.2019 at 10:29

22. Beer, D.; Redden, J.; Williamson, B.; Yuill, S. (2019). Landscape Summary: Online Targeting. CDEI. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819057/Landscape_Summary_-_Online_Targeting.pdf. Retrieved on 09.10.2019 at 10:31

23. https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation. Retrieved on 03.10.2019 at 12:48

24. Proudman, J. (2018). Speech: Cyborg supervision – the application of advanced analytics in prudential supervision. Bank of England. Workshop on Research on Bank Supervision. https://www.bankofengland.co.uk/-/media/boe/files/speech/2018/cyborg-supervision-speech-by-james-proudman. Retrieved on 15.10.2019 at 17:11

25. Carney, M. (2019). Speech: A Platform for Innovation. Bank of England. Remarks at Innovate Finance Global Summit, London. https://www.bankofengland.co.uk/-/media/boe/files/speech/2019/a-platform-for-innovation-remarks-by-mark-carney. Retrieved on 15.10.2018 at 17:15

26. Carney, M. (2017). Speech: The Promise of FinTech – Something New Under the Sun?. Deutsche Bundesbank G20 Conference on 'Digitising Finance, Financial Inclusion and Financial Literacy', Wiesbaden. https://www.bankofengland.co.uk/-/media/boe/files/speech/2017/the-promise-of-fintech-something-new-under-the-sun.pdf?la=en&hash=0C2E1BBF1AA5CE510BD5DF40EB5D1711E4DC560F. Retrieved on 16.10.2019 at 12:08

27. Proudman, J.. (2019). Speech: Managing Machines – the Governance of Artificial Intelligence. Bank of England. FCA Conference on Governance in Banking. https://www.bankofengland.co.uk/-/media/boe/files/speech/2019/managing-machines-the-governance-of-artificial-intelligence-speech-by-james-proudman. Retrieved on 16.10.2019 at 11:57

28. Bracke, Ph. et al. (2019). Machine Learning Explainability in Finance: an Application to Default Risk Analysis. Bank of England. Staff Working Paper No816. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf. Retrieved on 16.10.2019 at 12:00

29. Bank of England; Financial Conduct Authority. (2019). Machine Learning in UK Financial Services. https://www.fca.org.uk/publication/research/research-note-on-machine-learning-in-uk-financial-services.pdf. Retrieved on 18.10.2019 at 15:43

30. Anupam Datta, Shayak Sen, and Yair Zick. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In Proceedings of IEEE Symposium on Security & Privacy 2016, pages 598-617, 2016.

31. https://www.bankofengland.co.uk/about/governance-and-funding. Retrieved on 15.10.2019 at 16:54.

32. https://www.bankofengland.co.uk/knowledgebank/what-is-the-prudential-regulation-authority-pra. Retrieved on 15.10.2019 at 16:59

33. Financial Conduct Authority. (2017). Our Mission 2017: How We Regulate Financial Services. https://www.fca.org.uk/publication/corporate/our-mission-2017.pdf. Retrieved on 23.10.2019 at 10:25

34. Falk, M. (2019). Artificial Intelligence in the Boardroom. FCA Insight. https://www.fca.org.uk/insight/artificial-intelligence-boardroom. Retrieved on 23.10.2019 at 10.25

35. Woolard, C. (2019). The Future of Regulation: AI for Consumer Good. Speech by Christopher Woolard, Executive Director of Strategy and Competition at the FCA, delivered at The Alan Turing Institute's AI ethics in the financial sector conference. https://www.fca.org.uk/news/speeches/future-regulation-ai-consumer-good. Retrieved on 23.10.2019 at 10:34

36. Croxson, K.; Bracke, P.; Jung, C. (2019). Explaining Why the Computer Says 'No'. FCA Insight. https://www.fca.org.uk/insight/explaining-why-computer-says-no. Retrieved on 23.10.2019 at 10:38

37. Financial Conduct Authority. (2019). FCA Research Agenda: April 2019. https://www.fca.org.uk/publication/corporate/fca-research-agenda.pdf. Retrieved on 23.10.2019 at 10:41

38. Hunt, S. (2017). From Maps to Apps: the Power of Machine Learning and Artificial Intelligence for Regulators. Speech: Beesley Lecture Series on Regulatory Economics. 19 October 2017. https://fca.org.uk/publication/documents/from-maps-to-apps.pdf. Retrieved on 23.10.2019

39. Dungate, J. (2019). New Collaboration with the FCA on Ethical and Regulatory Issues Concerning the Use of AI in the Financial Sector. The Alan Turing Institute. https://www.turing.ac.uk/news/new-collaboration-fca-ethical-and-regulatory-issues-concerning-use-ai-financial-sector. Retrieved on 23.10.2019 at 10:51

40. Woolard, C. (2019). Regulation in a Changing World. Speech by Christopher Woolard, Executive Director of Strategy and Competition at the FCA, delivered at the City of London/ Cicero event on Future of Regulation. https://www.fca.org.uk/news/speeches/regulation-changing-world. Retrieved on 23.10.2019 at 10:54

41. Bank of England; Financial Conduct Authority. (2019). Machine Learning in UK Financial Services. https://www.fca.org.uk/publication/research/research-note-on-machine-learning-in-uk-financial-services.pdf. Retrieved on 18.10.2019 at 15:43

42. AI HLEG. (2019). Definition of AI: Main Capabilities and Disciplines. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines. Retrieved on 24.10.2019 at 12:52

43. AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Retrieved on 24.10.2019 at 12:54

44. AI HLEG. (2019). Policy and Investment Recommendations for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence. Retrieved on 24.10.2019 at 12:56

45. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-Intelligence. Retrieved on 24.10.2019 at 12:41

46. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf. Retrieved on 30.10.2019 at 16:41

47. Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/. Retrieved on 05.11.2019 at 14:26

48. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. https://arxiv.org/pdf/1602.04938.pdf. Retrieved on 05.11.2019 at 14:37

49. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. (2018). Anchors: High-Precision Model-Agnostic Explanations. https://homes.cs.washington.edu/~marcotcr/aaai18.pdf. Retrieved on 05.11.2019 at 14:40

50. Datta, A.; Sen, S.; Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf. Retrieved on 05.11.2019 at 15:53

51. Lundberg, Scott M., and Su-In Lee. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems

52. Bengio, Y. et al. (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. https://arxiv.org/pdf/1502.03044.pdf. Retrieved on 05.11.2019 at 15:56

53. Lei, T.; Barzilay, R.; Jaakkola, T. (2016). Rationalising Neural Predictions. Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology. https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf. Retrieved on 05.11.2019 at 15:58

54. Feldman, M. et al. (2015). Certifying and Removing Disparate Impact. arXiv:1412.3756v3 [stat.ML] 16 Jul 2015

55. Redmon, J; Divvala, S.; Girshick, R; Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640v5 [cs.CV] 9 May 2016. https://arxiv.org/pdf/1506.02640.pdf. Retrieved on 03.01.2020 at 10:01

56. Yang, Y.; Cer, D.; Ahmad, A.; Guo, M.; Law, J.; Constant, N.; Abrego, G.H.; Yuan, S.; Tar, Ch.; Sung, Y.; Strope, B.; Kurzweil, R. (2019). Multilingual Universal Sentence Encoder for Semantic Retrieval. arXiv:1907.04307v1 [cs.CL] 9 Jul 2019. https://arxiv.org/pdf/1907.04307.pdf. Accessed on 03.01.2020 at 10:04

57. Goodfellow, I.; Bengio, Y.; Courville, A. (2016). Deep Learning. MIT Press

58. Brainard, L. (2018). What Are We Learning about Artificial Intelligence in Financial Services. Speech at Fintech and the New Financial Landscape, Philadelphia, Pennsylvania. Federal Reserve. https://www.federalreserve.gov/newsevents/speech/brainard20181113a.htm. Retrieved on 03.01.2020 at 12:41

59. Hardt, M.; Price, E.; Srebro, N. (2016). Equality of Opportunity in Supervised Learning. arXiv:1610.02413v1 [cs. LG] 7 Oct 2016. https://arxiv.org/pdf/1610.02413.pdf. Retrieved on 06.01.2020 at 12:54

60. Sutton, R.; Barto, A. (2018). Reinforcement Learning: An Introduction. Complete draft. The MIT Press

61. Zhang, S.; Sutton, R. (2018). A Deeper Look at Experience Replay. arXiv:1712.01275v3 [cs.LG] 30 Apr 2018. https://arxiv.org/pdf/1712.01275.pdf.Retrieved on 06.01.2020 at 14:29

62. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. (2014). Generative Adversarial Nets. arXiv:1406.2661v1 [stat.ML] 10 Jun 2014. https://arxiv.org/pdf/1406.2661.pdf. Retrieved on 07.01.2020 at 11:48

63. Graves, A. (2014). Generating Sequences With Recurrent Neural Networks. arXiv:1308.0850v5 [cs.NE] 5 Jun 2014. https://arxiv.org/pdf/1308.0850.pdf. Retrieved on 07.01.2029 at 12:44

64. Cassells, K. (2019). Is third party insurance cheaper than fully comprehensive insurance? https://www.uswitch.com/car-insurance/is-third-party-insurance-cheaper-than-comprehensive-insurance/

65. European Commission. (2020). On Artificial Intelligence – A European approach to excellence and trust. White paper. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Retrieved on 02.03.2020 at 07:11

66. Information Commissioner's Office. (2020). Guidance on the AI auditing framework: Draft guidance for consultation. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf. Retrieved on 02.03.2020 at 07:30

Contacts

Matthew Dodgson
Director

M: +44 (0)7801 766052
E: matthew.dodgson@pwc.com

Fabrice Ciais
Associate Director

M: +44 (0)7843 334241
E: fabrice.ciais@pwc.com

Kiril D Georgiev
Manager

M: +44 (0)7710 036144
E: kiril.d.georgiev@pwc.com

RITM3009303