Response to the Bank of England Discussion Paper in Artificial Intelligence and Machine Learning (DP5/22)

The following is a response submitted to the Bank of England Discussion Paper in Artificial Intelligence and Machine Learning (DP5/22).
www.bankofengland.co.uk/prudential-regulation/publication/2022/october/artificial-intelligence

We welcome this discussion paper since an effective but agile regulatory regime that enables rapid innovation in the application of AI while protecting financial institutions, customers, and data subjects is key to securing the UK's future as a global leader in financial services technology.

This response comes from the perspective of a technologist who has been involved in the development of AI systems in the regtech sector that are now used by thousands of financial institutions globally, including Tier 1 banks in the UK and overseas. I have also developed AI governance frameworks that have enabled processes that use these solutions to pass review by the UK FCA, US Federal Reserve, HK Monetary Authority, and Tier 1 banks, and also used in the deployment of AI to manage critical services at Google Inc.

Joe Faith
shiftingrock.com
20th October 2022

Supervisory authorities' objectives and remits

Q1: Would a sectoral regulatory definition of AI, included in the supervisory authorities' rulebooks to underpin specific rules and regulatory requirements, help UK financial services firms adopt AI safely and responsibly? If so, what should the definition be?

There is a danger that any definition of AI will be superseded by technology; and it also slightly misses the point about the impact of AI. What matters about AI, from the regulatory and safety point of view, is not the nature of the technology itself, but that judgements that may impact financial institutions, customers, or data subjects, are delegated to computer processes. The technology that implements the judgement could be as simple as an 'if…then' statement. What matters is not the technology involved but that a computer is making a judgement that may impact humans or institutions.

'Judgement' in this sense is used specifically to refer to decisions where the output cannot be defined in terms of what computer scientists refer to as an 'effective procedure' -- i.e. which is always guaranteed to produce a correct answer. For example simple accounting or arithmetic operations are effective procedures since their correctness can be definitively

determined. Cases where there are no such guarantees can be considered as requiring some level of judgement, and it is the involvement of computer processes in these cases that require regulatory oversight, regardless of the sophistication of the specific technology involved.

Q2: Are there equally effective approaches to support the safe and responsible adoption of AI that do not rely on a definition? If so, what are they and which approaches are most suitable for UK financial services?

I have found the US Federal Reserve Guidance on Model Risk Management to be a useful framework for managing model and AI risk in general, even though it was originally targeted at the risk posed by the use of models in portfolio management and other financial applications. Their guidance includes a definition of 'model' as 'a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates'. The specific inclusion of the term 'estimate' in this definition corresponds closely to the definition of 'judgement' used above.

See https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm, and the responses to Q10, 11, 17 and 18 below.

Benefits, risks, and harms of AI

Q3: Which potential benefits and risks should supervisory authorities prioritise?

As with the definition of AI, any prioritisation of risks is likely to be superseded as soon as it is determined. Any risk prioritisation may also have the unintended consequence of relieving financial institutions of the responsibility of determining the specific risks inherent in their own AI, if they feel that those risks prioritised by supervisory authorities do not apply to them.

It is preferable to require *all* financial institutions to take a risk management approach in which they are required to identify, analyse, and estimate the risks associated with their own specific applications, processes, and systems.

Q4: How are the benefits and risks likely to change as the technology evolves?

One common risk as the technology evolves that has received little attention is that of 'de facto' black boxes.

There has been much discussion of how machine learning in general, and deep learning in particular, can produce black box solutions for which it is not intelligible how or why particular decisions are made. But a possibly larger problem is that of the adoption of technologies that may be made intelligible in theory, but not in practice by the engineers and data scientists who adopt them.

This is because many technologies that were once state of the art, and only used by the individuals and institutions that developed them and who were familiar with their limitations and suitable applications, are now freely available as packages for re-use. For example, language transformer models, such as BERT or GPT, that require substantial resources to train, are now freely available. This has accelerated adoption of these technologies, since it is now quick and easy for even inexperienced data scientists to use them. But this has the unintended consequence that these users may have little understanding of how they work or how they may fail. These packages are not strict 'black boxes' since they may be open source, well documented, and with clear decision criteria that are understood by their original developers. But they are black boxes in practice, since those factors are not well understood by the developers who are now using them.

BERT and GPT, for example, perform well on the text generation tasks they were originally developed for, and as such are suitable for use in low risk applications such as copywriting. But these models are now being used, via transfer learning, on tasks such as named entity recognition in the use of financial risk detection: an application that carries higher risk and for which they are not necessarily suitable.

Q5: Are there any novel challenges specific to the use of AI within financial services that are not covered in this DP?
Q6: How could the use of AI impact groups sharing protected characteristics? Also, how can any such impacts be mitigated by either firms and/or the supervisory authorities?

A particular risk for the impact of AI on groups with protected characteristics is the issue of 'hidden stratification'.

This is a recognised problem in the application of machine learning to medical diagnoses. For example, an algorithm for diagnosing tumours may be very accurate on the majority of tumours that happen to be benign, but have low accuracy on the small number of malignant tumours. Machine learning training regimes will typically optimise for overall accuracy across the entire sample population, but this may result in sacrificing accuracy on the minority of higher risk cases in order to increase accuracy on the more common, though lower risk, cases. The solution, in the case of medical diagnoses, is to 'stratify' the training and testing samples, to ensure performance on high risk classes can be measured separately from aggregate diagnostic performance. See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7665161/

An example of this problem in the case of the application of AI to the financial sector is that one particular ethnic group may be classed as high risk, and so denied access to financial services, because there is insufficient training data to enable a machine learning regime to develop a more accurate assessment of risk.

Q7: What metrics are most relevant when assessing the benefits and risks of AI in financial services, including as part of an approach that focuses on outcomes?

As discussed in the answer to Q10 and Q11 below, I would advocate a risk-based approach to AI governance. This involves identifying and analysing the risks involved in the adoption of an AI, including their likelihood and impact. The specific metrics for measuring these will depend on the nature of the application, and could vary from the denial of access to financial services to specific groups of individuals, to the risks of exposure to financial crime.

Rather than trying to weigh the risks against the benefits, the investment question should be whether the business benefits justify the investment, taking into account the effort and expense required to mitigate the risks to an acceptable level.

Regulation

Q8: Are there any other legal requirements or guidance that you consider to be relevant to AI?

--

Q9: Are there any regulatory barriers to the safe and responsible adoption of AI in UK financial services that the supervisory authorities should be aware of, particularly in relation to rules and guidance for which the supervisory authorities have primary responsibility?

--

Q10: How could current regulation be clarified with respect to AI?
Q11: How could current regulation be simplified, strengthened and/or extended to better encompass AI and address potential risks and harms?

I would advocate mandating the adoption of a risk management approach to AI Governance for all financial institutions.

I have developed AI governance policies for multiple tech and fintech companies that enabled them to make effective use of AI and machine learning in the face of ill-defined, and rapidly changing regulatory frameworks and stakeholder concerns. The resulting products and processes have been reviewed and approved by Tier 1 financial institutions, by regulatory bodies in the UK, US, and Hong Kong, and by the owners of some of the highest revenue services in the industry. See https://shiftingrock.com/cases.html for examples, and I would be happy to share more details of these policies and risk assessments on request.

In each case I took a risk management approach. This involves identifying and analysing the particular risks of each system and application, estimating their likelihood and impact, and defining, prioritising, and implementing mitigations where appropriate. Typical mitigations include: gathering and labelling large data sets for training and testing the solution, including for high-risk minority classes; ensuring the decisions of the solution can be effectively explained to, and reviewed by, a human domain expert; human-in-the-loop QA of a proportion of decisions; ongoing monitoring of inputs to ensure they fall within the distributions encountered during development, and resampling and retraining where they do not, etc.

However the reason for the risk-based approach is that there is no one-size-fits-all checklist of risks -- or mitigations -- that fit every application or technology. In some cases risks may be judged to be low impact or low likelihood, in which case they may not require resolution at all. Others may require ongoing monitoring. Others may be so severe they require de-risking even at the proof of concept stage.

What is vital is that the (1) developers of the solution undertake a risk assessment, and (2) there is effective review and challenge of this assessment by a 'risk steward' who has the information and opportunity to understand and challenge that risk assessment. Where solutions are developed externally to financial institutions, then vendors should be prepared to share the results of their internal risk assessments with customers.

A risk-based governance approach has two other advantages, in addition to enabling effective oversight. The first is that it has low overhead -- or rather the overhead of mitigation is proportional to the risk. Not all AI systems carry significant risk; a risk-based approach enables the mitigation effort to be focussed on the most severe. This factor is important since it avoids governance becoming a brake on development velocity. In some cases I have seen effective governance can *accelerate* development velocity, since development effort can be focussed on the risks that matter, while the mitigation of low impact or low likelihood risks are de-prioritised.

The second advantage is that it empowers those who are most knowledgeable of the possible failure modes and risks of the system -- the developers themselves -- to undertake their own assessment, while still requiring them to justify that assessment to an independent party. Rather than create an adversarial regime in which, possibly irrelevant, requirements are imposed from without, a risk-management approach invites those best placed to solve the problem to suggest their own solutions.

Q12: Are existing firm governance structures sufficient to encompass AI, and if not, how could they be changed or adapted?
Q13: Could creating a new Prescribed Responsibility for AI to be allocated to a Senior Management Function (SMF) be helpful to enhancing effective governance of AI, and why?

As suggested in the responses to Q10 and Q11, effective governance of AI requires enabling effective challenge by an individual who is independent of development or procurement of the AI systems, but has the information and opportunity to review the risk assessment and risk mitigation for those systems. I believe this requires an allocated SMF, similar to that of a Data Protection Officer or Information Security Officer.

Q14: Would further guidance on how to interpret the 'reasonable steps' element of the SM&CR in an AI context be helpful?
--

Q15: Are there any components of data regulation that are not sufficient to identify, manage, monitor and control the risks associated with AI models? Would there be value in

a unified approach to data governance and/or risk management or improvements to the supervisory authorities' data definitions or taxonomies?
--

Q16: In relation to the risks identified in Chapter 3, is there more that the supervisory authorities can do to promote safe and beneficial innovation in AI?
--

Q17: Which existing industry standards (if any) are useful when developing, deploying, and/or using AI? Could any particular standards support the safe and responsible adoption of AI in UK financial services?
Q18: Are there approaches to AI regulation elsewhere or elements of approaches elsewhere that you think would be worth replicating in the UK to support the supervisory authorities' objectives?

As discussed above, I have found the US Federal Reserve Guidance on Model Risk Management to be an effective framework for managing model and AI risk in general. Although it was originally written to address the risk posed by the use of models in portfolio management and other financial applications, two aspects in particular are particularly salient to the problem of AI governance. The first is to frame the problem of governance as a risk management problem, rather than presenting a checklist of requirements for developers to meet. The second is to insist on effective challenge and oversight at the highest level:

> "Strong governance … includes documentation of model development and validation that is sufficiently detailed to allow parties unfamiliar with a model to understand how the model operates, as well as its limitations and key assumptions ... Model risk governance is provided at the highest level by the board of directors and senior management when they establish an organization-wide approach to model risk management. Board members should ensure that the level of model risk is within their tolerance. A banking organization's internal audit function should assess the overall effectiveness of the model risk management framework, including the framework's ability to address both types of model risk for individual models and in the aggregate. "
> https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Q19: Are there any specific elements or approaches to apply or avoid to facilitate effective competition in the UK financial services sector?

I would hope to avoid a 'checklist' approach to the governance of AI, since I do not believe there is a one-size-fits-all solution to all technologies and applications.

For example, a common requirement proposed is that all AI decisions should be transparent and explainable. Although I would agree that this is a desirable characteristic in general, in some cases it is a 'nice to have'.

For example, a common use of AI, broadly understood, is in entity resolution or search relevancy. This is deciding whether a source of information about named individuals (such as a credit report, adverse media reports, or reports on disqualified directors) applies to the

individual that the financial institution is dealing with, when they only have a name in common. Although this may seem like a simple problem, AI is increasingly required to use contextual clues, such as locations, ages, or employment histories, to determine the identity of individuals.

In some high risk applications, such as enhanced due diligence processes, it may be necessary for a human reviewing the case to understand the reasoning why a source of information is judged to be relevant. But in other common cases -- such as searching for a disqualified director on the Companies House register -- it should be possible to provide a search function to a user without requiring full explanation of how the match score is derived. A better solution to the risk of mis-identification in these cases is to present the search results in descending order of match score, and make it possible for the human to review the source of information themselves.

If all applications of AI are required to fulfil a common checklist of requirements then the response of developers will be to under-report what constitutes AI, and significant risks may be missed. A better solution is to require developers to consider the risks in all systems that involve making judgements, as defined above, and to report and mitigate them appropriately.