Sarah Breeden: Engaging with the machine - AI and financial stability

Speech by Ms Sarah Breeden, Deputy Governor for Financial Stability of the Bank of England, at the HKMA-Bank for International Settlements (BIS) Joint Conference "Opportunities and Challenges of Emerging Technologies in the Financial Ecosystem", Hong Kong, 31 October 2024.

* * *

I'm going to speak today about what the increasing power and use of Artificial Intelligence (AI) – in particular generative AI[1] - might mean for financial stability, and how central banks and regulators should respond.[2]

As I'll go onto unpack, I start from the premise that generative AI models have distinctive features compared to other modelling technology. They can learn and evolve autonomously and at speed, based on a broad range of data, with outputs that aren't always interpretable or explainable and objectives that may be neither completely clear nor fully aligned with society's ultimate goals.

AI is expected to bring considerable potential benefits for productivity and growth in the financial sector and the rest of the economy. But for the financial sector to harness those benefits we, as financial regulators, must have policy frameworks that are designed to manage any risks to financial stability that come with them. Economic stability underpins growth and prosperity. It would be self-defeating to allow AI to undermine it.

Furthermore, while there is significant uncertainty about how far and how fast AI will be adopted, we don't want to be left in the position of choosing between, on the one hand, letting a powerful new technology threaten financial stability, and on the other, preventing its use and losing out on growth and innovation - simply because we don't have the policy frameworks to enable its safe adoption.

The financial services industry is in the early stages of adopting GenAI. But as we look ahead, I think we should have a watchful eye on two issues:

- First, at a microprudential level (where we seek to ensure the safety and soundness of individual firms), central banks and financial regulators should continue to assure themselves that technology-agnostic regulatory frameworks are sufficient to mitigate the financial stability risks from AI, as models become ever more powerful and adoption increases. We need to be focused in particular on ensuring that managers of financial firms are able to understand and manage what their AI models are doing as they evolve autonomously beneath their feet.
- Second, we should be alive to the possible need for macroprudential interventions to support the stability of the financial system as a whole. We should keep our regulatory perimeters under review, should the financial system become more dependent on shared AI technology and infrastructure systems. And our stress testing frameworks could usefully evolve in time to assess whether AI models used 'in the front line' of financial firms' businesses could interact with each other in ways that are hard to predict ex ante – for example, when used for trading,

could we see sophisticated forms of manipulation or more crowded trades in normal times that exacerbate market volatility in stress.

To be clear, I don't think that, at the Bank of England, we are yet at the point where we need to change our tech-agnostic microprudential approach or where macroprudential policy is needed. But the power and use of AI is growing fast, and we mustn't be complacent. We know from past experience with technological innovation in other sectors of the economy that it's hard retrospectively to address risks once usage reaches systemic scale.[3]

I also recognise that many of the issues surrounding AI could have broader implications, which governments will decide how best to manage across the economy as a whole. But given our responsibility for financial stability, we need to consider what, if anything, might be needed in the financial system in advance of any broader government action.

To that end, we are launching an AI Consortium of the private sector and AI experts to help us understand more deeply not only AI's potential benefits but also the different approaches firms are taking to managing those risks which could amount to financial stability risks.[4] We will consider what we can do to spread best practices widely in the industry and whether further regulatory guidelines and guardrails are needed. And our Financial Policy Committee (FPC) will publish its assessment of AI's impact on financial stability and set out how it will monitor the evolution of those risks going forward. As we do so, we will work with the UK Financial Conduct Authority (FCA), the government and international counterparts – to support AI's safe adoption as the best contribution we can make to harnessing its benefits for growth.

The use of AI in financial services

The context for my remarks today is of course that AI is being used for a wide and increasing range of applications in financial services.

For the past five years, the Bank of England and the FCA have been running a periodic survey of how financial services firms in the UK are using AI and machine learning.[5] The latest one, earlier this year covered nearly 120 firms, including banks, insurers, asset managers, non-bank lenders and financial market infrastructures. We'll publish the full results soon.

But to preview some of the headlines, we've found that 75% of the firms surveyed are already using some form of AI in their operations, including all of the large UK and international banks, insurers and asset managers that responded. That's up from 53% in 2022.

17% of all use cases are using foundation models – models, including large language models like OpenAI's GPT4, which apply advanced machine learning (so-called deep learning) to very large quantities of data such that they can be applied across a wide range of use cases.

Some of the most prevalent early use cases for AI have been fairly low risk from a financial stability standpoint. 41% of respondents are using AI to optimise internal processes, while 26% are using AI to enhance customer support, helping to improve efficiency and productivity.

But many firms are also using AI to mitigate the external risks they face from cyber-attack (37%), fraud (33%) and money laundering (20%). For example, payment systems have long used machine learning automatically to block suspicious payments – and one card scheme is this year upgrading its fraud detection system using a foundation model trained on a purported one trillion data points.

Potentially more significant use cases from a financial stability perspective are emerging. 16% of respondents are using AI for credit risk assessment, and a further 19% are planning to do so over the next three years. Meanwhile, 11% are using it for algorithmic trading, with a further 9% planning to do so in the next three years. And 4% of firms are already using AI for capital management, and a further 10% are planning to use it in the next three years.

What makes AI different to previous modelling technology?

I want to spend some time now on five features which combine to make AI models warrant particular consideration, and why I think that might matter for financial stability.

First, AI models can be dynamic. The ways they turn input data into output results update automatically as they learn from new data. So, the way they behave can over time become misaligned with the original intention.

In the context of the financial system, AI models used for trading, with an objective function to make money, could evolve in a way that pursues that goal by learning the value of actively amplifying an external shock to market prices, or of profitably and autonomously colluding with another AI model.

The second particular challenge of AI models is their potential lack of explainability. These are typically complex models, with relationships between inputs and outputs that are constantly evolving as I've just described – and all this is taking place at considerable processing speed.[6]

Again, in the context of the financial system, this presents challenges. If an AI model classifies certain transactions as fraudulent or low-risk, or certain potential borrowers as a good or a bad credit - in ways that we can't easily explain - how do we know if that's an error with the model or the finding of an important pattern in the data that traditional analysis can't spot? How do we interrogate if an AI model telling us that sub-prime mortgages are low-risk is erroneous or not?

The third particular challenge of the latest AI models is the breadth of data on which they're trained, particularly for foundation models that are learning from huge numbers of large datasets from different sources, at a totally different scale to traditional models.

At that scale, and in light of the explainability challenges I've just talked about, knowing whether we are introducing misspecification into AI models through low-quality or biased training data is clearly extremely challenging.

Fourth, foundation models seem particularly prone to users coalescing around a small number of common models. The intellectual capital, computing power, and data needs to design and run such models make it a very expensive endeavour – and so their development seems likely to tend towards oligopoly. Indeed, in our survey this year, 44% of the third party AI models used by firms were from the top three model providers, compared to 18% in the previous survey conducted in 2022 - before the launch of ChatGPT accelerated interest in foundation models and generative AI.

It seems to me plausible that we could see widespread use of common foundation models, upon which downstream applications are dependent, not just across the financial system, but across the economy, and across borders. This introduces macro fragilities: an incident with a base model or a cloud provider supporting it could have systemwide implications. Or common models could increase the risks of correlated responses by market participants to shocks which amplify stress.

Fifth and finally, AI models are autonomous – adjusting automatically how they convert inputs to outputs, and in the case of generative AI, producing rich outputs in the form of text, images and video. That means the models could, in theory, be used to determine outcomes and make decisions without a senior manager that sufficiently understands the rationale for those decisions and is directly accountable for them. More than half (55%) of AI use cases in our latest survey have some degree of automated decision making, with a roughly 50:50 split between semi-autonomous decision-making (where there's human intervention at some point in the decision-making loop) and cases where the decision-making is completely automatic. That clearly poses challenges for financial firms' management and governance, and for supervisors. While the way a person makes decisions based on information can also be complex, opaque and hard to explain, a person can be held to account.

What might AI mean for microprudential supervision?

Despite these unique features, financial regulators (including in the UK) have tended to adopt a technology-agnostic approach in the way we seek to address its risks. In other words, we've expected firms to meet our existing rules on data management, model risk management, governance and operational resilience (including reliance on third parties), regardless of the technology they're using.

That is of course a sensible place to start. A tech-agnostic approach future proofs regulatory frameworks, by focusing on what matters (the outcomes) and not requiring perfect foresight on the part of regulators for how technology will evolve to deliver them.

And to date, a tech-agnostic approach seems to have worked pretty well for AI and machine learning. Firms have been quite cautious in the way they have deployed AI in their operations. In our engagement with them, including a Discussion Paper we published with the UK Financial Conduct Authority in 2022, many respondents thought there were no regulatory barriers to the safe and responsible of AI in the UK. That is good news. It means our regulation is not impeding the growth and productivity benefits of the technology that I mentioned earlier.[7]

The question we have to keep asking ourselves though, as AI models get ever more powerful and ever more widely adopted in a wider range of use cases, is whether we can continue to rely on existing regulatory frameworks - as these were not built to contemplate autonomous, evolving models with potential for decision making capabilities. Does a tech-agnostic approach continue sufficiently to mitigate the risks, or does AI necessitate a somewhat different approach? And given that AI is used across the economy and not just in finance, how might our expectations for development of models deployed in finance sit alongside those for models deployed in other industries?

That's why we're continuing our work on AI - let me highlight a few areas which we are particularly keen to explore.

On model risk management, are existing regulatory and supervisory frameworks sufficient to ensure firms understand what their AI models are doing as they evolve autonomously, now and in future, and are they able to constrain it where necessary? Respondents to our Discussion Paper highlighted the risk that the model risk management principles we set out[8] for banks last year (covering quantitative modelling in general, not just AI) might not be sufficient to ensure model users fully understand the third party AI models they deploy within their firms. Limited explainability of AI models is a particular focus. And so as regulators, we need further to consider what explainability means in the context of generative AI, what controls we should expect firms to have and what that means for our regulatory and supervisory frameworks.

Feedback to our Discussion Paper also noted the lack of clear, widely applicable standards around the data which AI models are trained on. Can we do more to ensure that firms are training AI models on high-quality, unbiased input data; that they can to a reasonable degree trace through how the model's behaviour is responding to changes in particular aspects of that training data; and that they can understand where the model is particularly dependent on certain segments of training data?

Finally on governance, a striking finding in our latest survey is that only a third of respondents describe themselves as having a complete understanding of the AI technologies they had implemented in their firms. Of course, at one level it's not surprising that this isn't at 100% - this is a fast-evolving technology, and there's some element of learning by doing. That said, as firms increasingly consider use of AI in higher impact areas of their businesses such as credit risk assessment, capital management and algorithmic trading, we should expect a stronger, more rigorous degree of oversight and challenge by their management and Boards – in particular given AI's autonomy, dynamism and lack of explainability.

Most respondents to our Discussion Paper agreed that practical guidance would be helpful on what 'reasonable steps' senior management might be expected to have taken with respect to AI systems to comply with regulatory requirements. Existing guidance is based on a time when autonomous decision-making technology such as AI was not widespread.

The Discussion Paper also raised the question of whether regulatory expectations for senior managers at firms – the Senior Managers and Certification Regime – should allocate a specific responsibility for AI, to create an incentive for meaningful accountability for AI deployment and oversight within firms. While most respondents were wary of this, we should continue to examine ways to enhance effective governance of AI – including to think about where we might be content for AI models to make automated decisions and where (and to what degree) there should be a human in the loop.

What might AI mean for macroprudential policy?

I've spoken so far about the microprudential aspects of AI – how could individual firms' use of AI pose risks to their safety and soundness, and through that financial stability.

But even if microprudential risks are managed well by individual firms, AI could pose risks to financial stability if it fails to take into account the impact of its actions on the rest of the financial system. Keeping the financial system safe is the focus of macroprudential policymakers such as the Bank of England's FPC. Indeed, the FPC will publish early next year its assessment of AI's impact on financial stability and set out how it will monitor the evolution of those risks.

An issue we worry about all the time as macroprudential policymakers is interconnectedness – where the actions of one institution can affect others, firms can become critical nodes, and firms can be exposed to common weaknesses. AI could both increase such interconnectedness and increase the probability that existing levels of interconnectedness threaten financial stability.

I've already talked a bit about how use of foundation AI models could tend towards reliance by the financial sector on shared AI technology and infrastructure systems. If use of AI models becomes more ubiquitous in finance, those could in turn depend on a small number of providers for data storage, model computation and deployment, and a small number of data aggregators for training data. Disruptions to these service providers could result in AI models that rely on their infrastructure becoming unavailable or performing poorly.

AI could also increase the probability of existing interconnectedness turning into financial stability risk – in particular through cyber-attacks. AI could of course improve the cyber defence capabilities of critical nodes in the financial system. But it could also aid the attackers – for example through deepfakes created by generative AI to increase the sophistication of phishing attacks.

There are other channels through which AI can have systemic risk consequences, particularly if they come to be used more in trading. For example, as noted in the IMF's recent Global Financial Stability ReportOpens in a new window, AI could lead to increased market speed and volatility under stress.

Specifically, multiple market participants using the same AI models and relying on a small number of AI service providers for trading could result in increasingly correlated trading behaviour. Particularly where such crowded trades are funded through leverage, a shock which causes losses for such trading strategies could be amplified into more serious market stress through feedback loops of forced selling and adverse price moves.

Indeed, some AI trading models might respond to such a scenario by seeking to exploit vulnerabilities in the trading algorithms and strategies of other firms in a manner which is individually rational but has adverse consequences for the overall financial system, by triggering or amplifying price movements in a manner which is destabilising for financial stability.

There is also the potential for system-wide conduct risk. If AI determines outcomes and makes decisions, what would be the consequences if, after a few years, such outcomes and decisions were legally challenged, with mass redress needed?

We have some tools in our macroprudential arsenal that we can use to address these issues.

Indeed, reliance on common technology providers, is already captured in the UK as the FPC recommended in 2021 the creation of a regime for critical third party service providers (CTP). That regime came into force last year, allowing a small number of third parties that provide services that are material to multiple firms and which are difficult to substitute easily or quickly to be designated by the UK Treasury for direct oversight by financial regulators. The regime was motivated by a more general concern than AI specifically, and recognised the limits as to how much an individual financial firm can accomplish through managing its own third party relationships. The Bank, the Prudential Regulation Authority (PRA) and the FCA has consulted on the rules that will apply to those third parties and is considering which third parties to recommend to the Treasury for designation.

Nevertheless, the CTP regime is designed to address the risk of failure or operational disruption at a critical node. AI could lead to a different kind of reliance, since those firms would be expected themselves to ensure that the third party model meets the same standards for model risk and data risk management as if it had been developed in-house. For firms using the most complex foundation models developed by third parties for material use cases in their businesses, that might be challenging to do in practice without visibility over how the model is designed and the capability to interrogate it. In due course, depending on how financial firms' use of AI evolves (particularly if it starts to be used in a material fashion for trading or core risk assessment), we may need to think again about the adequacy of the regulatory perimeter and whether some requirements applying directly to model providers themselves might be necessary.

The other tool which we can point at the financial stability vulnerabilities from AI is stress testing. We could perhaps use stress tests to understand how AI models used for trading whether by banks or non-banks could interact with each other. We could look to better understand reaction functions; seek to identify where elements of objective functions might cause them to evolve in ways which actively amplify shocks and undermine financial stability; and use these results to inform where intervention is required.

What should we do internationally?

In addition to pursuing these issues domestically, we also need to explore them with our international peers.

The international regulatory community has had great success in understanding new issues and innovations together – climate, cloud, crypto and stablecoin to name just a few – where we have learned together, from each other, and developed shared principles for how to approach such cross-border challenges.

To date, at the government level, we have seen international collaboration on principal issues of AI safety, including through the Bletchley and Seoul summits.

Looking ahead, it will be crucial for international bodies and national authorities to continue collaborating. This will help ensure we have the capacity to monitor AI adoption across the global financial system, to assess whether our current regulatory frameworks adequately address vulnerabilities, and to consider ways to enhance those where necessary. Our cooperation can also strengthen our collective resilience should bad actors try to use AI to destabilise the financial system, including through cyber attacks. I welcome the work from international groups, including the FSB, IOSCO, and G7, to consider the implications of AI and to build our common understanding. Policy work might well be premature now, but this part of finance is moving quickly. We should continue to further our understanding of AI together so that if guidelines and guardrails are needed in future, we are ready.

Conclusion

AI models come with a unique and powerful combination of features. As I said, they learn and evolve autonomously and at speed, based on a broad range of data, with outputs that aren't always interpretable or explainable and objectives that may be neither completely clear nor aligned with society's goals.

Even experts don't agree on how far and how fast AI (including in finance) will evolve. So we need both to be humble and to be prepared. While not jumping to knee-jerk policy responses, we need to keep under review whether our microprudential and macroprudential policies remain sufficient to maintain financial stability. In so doing, we can harness AI's considerable benefits for economic growth in a safe and sustainable way.

I would like to thank Michael Yoganayagam for his assistance in drafting these remarks. I would also like to thank Andrew Bailey, Colette Bowe, Mohammed Gharbawi, Bernat Gual-Ricart, Amy Lee, Owen Lock, Harsh Mehta, Tom Mutton, Danny Walker, Mei Jie Wang, Ewa Ward and Sam Woods for helpful input.

[1] Generative AI is a subset of AI-machine learning technologies (AI/ML), distinguished by its ability to create new content, including understandable and meaningful text or human languages, based on the data it was trained on.

[2] This speech covers the financial stability implications of use of AI by financial firms. Of course, use of AI by central banks ourselves also presents many opportunities to enhance our own analysis and modelling, as well as helping to inform our policy work

on the use of AI in finance. My colleague James Benford spoke last month about how we're seeking effectively to deploy AI within the Bank of England: TRUSTED AI: Ethical, safe, and effective application of artificial intelligence at the Bank of England speech by James Benford | Bank of England.

[3] Londoners overwhelmingly against TfL decision to ban Uber, analysis of social media posts reveals | London Evening Standard | The StandardOpens in a new window

[4] Artificial Intelligence Consortium membership call for interest | Bank of England

[5] 2022: Machine learning in UK financial services | Bank of England. 2019: Machine learning in UK financial services | Bank of England.

[6] Monsters in the deep? – speech by Jonthan Hall, 7 May 2024. In this speech, it was noted that in 2014, researchers found that, with a tiny change to the pixel field, an image of a panda could be incorrectly classified by an AI model as a gibbon, even though it looked completely unchanged to humans. This emphasises the potential lack of explainability of AI models.

[7] FS2/23 – Artificial Intelligence and Machine Learning | Bank of England

[8] SS1/23 – Model risk management principles for banks | Bank of England