SUERF Policy Brief
No 1163, May 2025

How central banks can meet the financial stability challenges arising from artificial intelligence


Jón Danielsson | SUERF Fellow and Systemic Risk Centre, London School of Economics
Andreas Uthemann | Bank of Canada and Systemic Risk Centre, London School of Economics

Keywords: Artificial Intelligence, regulations, financial stability, central banks

JEL codes: C02, C52, D50, G20, G28

Abstract

The growing use of artificial intelligence (AI) poses difficult challenges for the financial authorities. AI allows private-sector firms to optimise against existing regulatory frameworks, helps those who want to damage the financial system, amplifies wrong-way risk and speeds up financial crises. It also gives the authorities new tools for executing their mandate. The authorities could become more effective stewards of the financial system by gaining expertise in AI, setting up AI-to-AI links, developing triggered facilities and incorporating AI into their monitoring frameworks.

Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Bank of Canada. We thank Nikola Tarashev for valuable comments. All mistakes and opinions are ours.

Introduction

As the rapid adoption of artificial intelligence (AI) is transforming the financial system for the better, it also poses significant and poorly understood threats to financial stability. How the central banks opt to engage with AI will determine whether systemic financial risk increases or decreases.

It is tempting to consider the impact of AI as just another step in the long line of technological improvements, but we think that would be a mistake. Technological advancement has improved processes and gives better signals to decision-makers, but AI also makes decisions. In the nomenclature of Norvig and Russell (2021), AI is a "rational maximising agent", a notion that aligns with standard economic paradigms.

Would AI be better than the existing regulatory setup? Its superior ability to process information will undoubtedly be invaluable to the authorities. However, AI cannot detect the most serious threats to the system's stability because of the financial system's almost infinite complexity, inherent constraints and the nature of crises. The data it would need to identify the build-up of fragilities is not available until it is too late. The vulnerabilities that led to the global financial crisis in 2008 are an example. The necessary data was not collected deliberately, and much of the data that did exist, could not legally be aggregated and used for systemic risk analysis.

Our objective in this article is not to study AI technology and how it improves the provision of financial services, as one can find excellent material on that elsewhere. Instead, building on our existing work (Danielsson et al. 2023; Danielsson and Uthemann 2024), our emphasis is on how AI affects financial regulations and stability. We specifically examine how the financial authorities can effectively engage with AI, both to contain the risks arising from it and to harness it in the delivery of their mandate.

How AI can undermine financial stability

AI does not appear to pose fundamental threats to financial stability, but instead interacts with existing vulnerability channels. Of particular concern are malicious use, wrong-way risk, synchronisation and speed.

Malicious use and the defenders' dilemma

Technology has always helped improve the financial system, but has also provided new ways to destabilise it. Nathan Rothschild, for example, crashed the London stock market by using fast couriers – some say carrier pigeons – to spread false rumours of Napoleon's victory at Waterloo, and the sophisticated algorithmic trading systems of today cause flash crashes.

It is the same with AI. It helps criminal organisations to identify loopholes and manipulate systems for illegal profit, and terrorist groups to orchestrate synchronised attacks on the financial infrastructure. Of particular concern is nation states recruiting AI for unconventional warfare by targeting vulnerabilities in their adversaries' financial systems while maintaining plausible deniability.

The growing use of AI poses particular challenges to those tasked with defending the financial system against attacks – what we term the "defender's dilemma". The attackers need only to find a single vulnerability, whereas the defenders must monitor the entire system against potential attacks. The attackers consequently require considerably fewer resources than the defenders; an asymmetry that gets worse with the increasing ability of AI systems.

As a rational maximising agent, AI can also, by itself, act maliciously. Scheurer et al. (2024) provide a good example: a Large Language Model (LLM) was instructed to comply with securities laws and maximise profits. The aim was to investigate how AI would cope with those two potentially conflicting objectives. When given private information, the LLM engaged in illegal insider trading and then lied about it to its human overseers. This exemplifies how AI gives rise

to new types of model risk simply because it does what it thinks it should, only to find unexpected and potentially harmful paths to attain the goals it is given.

Guarding against AI engaging in this kind of behaviour is not easy. A possible solution is to always have humans in-the-loop to make decisions based on AI recommendations, or humans on-the-loop to supervise AI. But it is not as straightforward as it might seem. Human experts cannot effectively oversee AI that is operating at machine speeds, and in highly competitive markets, those who take humans out-of-the-loop gain the upper hand. Furthermore, as discussed below, a human in/on-the-loop approach to crises interventions may make crises more likely and severe.

Consequently, it is naïve to assume that all we need for AI to optimize for the social good is to ensure we keep humans in-the-loop.

The existing regulatory structure may not be up to regulating a private-sector financial system that relies on AI. The supervisors have always faced difficult principal-agent problems when trying to align the interests of private-sector risk takers with those of society. Those problems get worse with AI. Conventional incentive mechanisms that work well with human agents, such as rewards and punishments, are ineffective with AI systems. The one-sided principal-agent problem becomes two-sided: principal-agent–AI.

Wrong-way risk

Wrong-way risk refers to a situation where the volume of risky activities increases in line with the riskiness of those activities. AI gives rise to particularly dangerous types of wrong-way risk because of how it builds trust. Using it for simple, repetitive and non-critical tasks plays to its strength. As AI systems demonstrate competence in increasingly complex assignments, we might end up with an AI version of the Peter principle, where AI is progressively trusted with more difficult tasks until its capabilities no longer match the job requirements.

The typical response to such an eventuality is to maintain that AI will not be given many responsibilities and that there will always be humans in/on-the-loop, except for the most innocuous tasks. That is not credible. In the highly competitive private financial sector, AI brings significant competitive benefits to those who harness it. These competitive pressures strongly oppose any attempts at constraining AI adoption.

The reason for AI wrong-way risk is that AI will not have all the information it needs to regulate the financial system effectively. Because unique circumstances that differ significantly from those of the past characterise every financial crisis, the data required for effective regulation does not exist until it is too late to do anything about the problem. In particular, we do not know how the public and private sectors will react to serious future stress. This remains an unknown because we cannot predict how future politicians, bank CEOs and central-bank governors will act.

AI excels at pattern recognition based on historical data, but struggles with unprecedented scenarios – the unknown unknowns at the heart of all crises. It could not be any other way. We learn from past crises, putting in place measures to prevent a repeat, so it is almost axiomatic that crises emerge where the supervisors are not patrolling. This is why AI will not have an intrinsic advantage over human macroprudential supervisors, and hence cannot solve the problem of financial crises.

Ultimately, this means that AI reliability is the lowest precisely when it is needed for the most important decisions – the AI wrong-way risk.

Procyclicality, market structure and synchronised behaviour

A central feature of financial markets is that many, even most, decisions are complementary: the optimal move of one market participant incentivises others to do the same. Economists call these strategic complementarities. This implies that market participants are often incentivised to synchronise their actions with one another, such as during speculative attacks, momentum trading and – especially – crises. AI significantly increases the systemic danger arising from strategic complementarities.

These complementarities are amplified when AI engines observe the decisions made by competing engines. Just like humans, AI engines train one another. The result is AI-to-AI communication channels that remain hidden until they encounter a particular situation, such as a shock. Then – during crises, for example – the actions of one engine significantly affect how others will react.

The potential for synchronised behaviour is directly affected by the strength of strategic complementarities and by how similarly market participants understand the world. When they have different views, they are more likely not to act in concert – some will sell and others will buy when a shock arises. In aggregate, this absorbs the shock. The more similar their information about the world becomes, the more likely market participants are to act as one, amplifying and deflating the same bubbles.

AI amplifies both drivers of synchronous behaviour, namely strategic complementarities and a similar world view. Let's start with the mathematical design of the neural networks that underpin AI. While competing AI will each have their own networks, the most successful engines are most likely to see widespread use and emulation. This limits variability. Even if the underlying network architecture and data for various engines are different, the training data is often similar, as is the object of neural network training – to maximise profit.

Furthermore, as described in Gambacorta and Shreeti (2025), the cost of designing and training state-of-the-art AI engines is significant and out of the reach of all but the largest and best-resourced organisations. Most private- and public-sector entities will have no choice but to get their AI engines from a handful of vendors. The consequence is risk monoculture, where a common view of risk drives risk takers to similar decisions, even when they have very heterogeneous objectives – pension funds invest differently to hedge funds, for example.

Market structure compounds the potential for synchronisation. Since only the largest private-sector institutions can afford to develop their own proprietary AI engines, the globally systemically important banks (the GSIBs) may end up reaping the strongest benefits. That advantage might be overcome by neo-banks, with their modern technology stacks and technologically attuned staff. The middle-tier institutions that already struggle with legacy systems, technical debt and staff accustomed to the pre-AI world find it difficult to compete with the GSIBs and neo-banks. This drives market concentration, entrenching the GSIBs and increasing systemic risk.

Taken together, the consequence is correlated behaviour that amplifies market booms and busts. In other words, AI is procyclical.

AI speed

The main immediate damage from financial crises arises from banks' instinct to protect themselves in times of stress. Danielsson (2024) terms this the one-in-a-thousand-day problem. When a shock arrives, banks need to decide whether to protect themselves or not – to stay or to run. They will act as one. All process the same information; all want to stay alive; all watch one another and the authorities. Getting it wrong has catastrophic consequences.

	Run	Stay
Crisis	✓Right decision	✗Wrong decision
No Crisis	✗Wrong decision	✓Right decision

If a bank perceives that a shock will not culminate in a crisis, it is optimal for it to stay, either by not reacting or by buying risky assets sold in panic. However, if a bank concludes that a shock will culminate in a crisis, it will run as fast

as it can by selling risky assets into a falling market and withdrawing liquidity. The first bank to run gets the best prices; the last faces bankruptcy. Speed is of the essence, which is why crises are so sudden and vicious. AI excels at rapidly processing large amounts of information; when it reaches its conclusion, it reacts quickly and decisively, speeding up crisis reactions.

When the AI engines collectively decide to stay, they stabilise the system, quickly absorbing the shock. In effect, they are doing the authorities' job for them.

When the engines decide to run, every AI wants to be the first to do so. Thus a crisis that previously took days or weeks to unfold will now happen in minutes or hours. AI destabilises the system and makes the job of the authorities harder.

AI lowers volatility and fattens tails by smoothing out small shocks and amplifying the large ones.

Implications for micro- and macroprudential regulations

The financial authorities face a fundamental challenge from private-sector AI. At the microlevel, supervision, at present founded on PDF reports, database dumps, periodic inspections and face-to-face conversations, will not keep pace with an AI-powered financial sector. The private-sector systems generate the required regulatory information in the format demanded by the authorities. Simultaneously, they optimise against those very requirements, creating an increasing asymmetry between the regulated and the regulators and undermining the effectiveness of the microprudential framework. Just one example is how a regulated entity can use AI to find economically equivalent forms of leverage that are treated differently under the leverage ratio, perhaps using derivatives contracts.

The macro-implications of AI are more serious. To begin with, the authorities might not detect these new risks because the immediate impact of AI on market dynamics will provide short-term stability at the expense of the increased potential for serious crises, as noted in the section on AI speed. The central banks might not notice because their systemic risk dashboards focus on the more visible aspects of financial markets. They might conclude that AI is unambiguously stabilising.

Even when authorities incorporate AI into their systemic risk analytics, they will find their stress response framework too slow. An example is that certain core Basel methodologies, such as the liquidity coverage ratio, might not provide the same protection in the future if the speed and efficiency of AI-controlled liquidity management undermine assumptions of run-offs. Major damage may have been done before the authorities have had time to respond.

Ultimately, if the financial authorities do not develop a credible response to AI, AI-induced financial crises become more likely. The reason is when a private-sector AI decides whether to stay or run, it will take the expected central-bank response into account. If it concludes that the public sector is insufficiently prepared, it may well be optimal for the private-sector AI to pre-emptively position itself for survival.

Policy responses

The policy authorities have to respond to AI. If they do so effectively, they can leverage it to perform their mandate better, as pointed out by the International Monetary Fund (2024). When they do not, they destabilise the financial system. Below, we outline the main areas where AI will probably have the strongest impact on the financial authorities and how they can respond best.

Data and federated learning

The financial authorities, especially those tasked with financial stability, find their job hampered by the lack of data sharing. In aggregate – and in principle – they have access to much relevant information; in practice, much of that data is inaccessible for those tasked with financial stability.

AI provides a valuable solution to the problem of data sharing, as it may not need access to the underlying supervisory-level data to be effective. Instead, the authorities can use federated learning to obtain the neural network weights. This means the actual training takes place on computer systems under the control of the "owners" of the data and only the weights are shared. The shared weights are then used in a common neural network that spans multiple authorities, allowing for a common understanding of financial stability. Federated learning has already proven very useful in anti-money laundering and fraud detection.

Since the weights in a typical over-parameterised neural network cannot be mapped onto the underlying training data, the sharing of weights is likely to be much less sensitive than the sharing of the underlying data.

Organisational structure

When AI first came on the central banks' radar, they often centred their response around divisions dealing with data, IT or innovation. However, given how AI can act as a powerful agent for systemic risk, financial stability should play a central role in AI policy. It would be best to integrate AI horizontally and vertically into the financial stability function. At the highest level, financial stability committees should include members with AI expertise, while at the operational level, AI specialists work alongside traditional financial stability experts. Ultimately, those responsible for system stability should have both the mandate and the expertise to address AI-related risks.

AI engines used by authorities

The authorities face a dilemma in developing their own AI capabilities. Open-source models run internally provide greater transparency and control, but are probably less capable than commercial alternatives. Models operated by commercial vendors (closed or open source) provide superior performance, but at the expense of security and sovereignty. The central banks will probably have no choice but to rely on outsourcing. Firms domiciled and operated in the same jurisdiction as the authority could then build their local expertise and address sovereignty and security concerns.

AI-to-AI links and benchmarking

AI provides an opportunity for authorities to interact more efficiently with regulated entities than is currently the case. This allows for the benchmarking of regulations and the quantification of systemic risk. Instead of relying on conventional reporting mechanisms, authorities can establish direct communication channels (API-to-API links) between their AI systems and those of regulated entities and other authorities. These API connections would allow regulatory AIs to query other engines about potential reactions to market developments or policy changes without requiring data sharing.

This will be of considerable value for microregulations as it will allow supervisors to check compliance, design regulations and benchmark private AI against regulatory standards.

The macro-authorities can employ such links to monitor the financial system in real-time, evaluate industry-wide feedback loops, test crisis-response scenarios and identify coordination risks before they materialise. These links also allow them to conduct counterfactual analysis, testing how private-sector AI would respond to various stress scenarios or policy interventions. This might include simulating market shocks, liquidity interactions or changes in regulatory parameters and then iteratively analysing the aggregate response across multiple institutions. In this way, the

authorities can identify coordination risks, dangerous feedback loops and tipping points before they manifest in actual markets. Authorities could, for example, determine what level of market stress might trigger synchronised selling across AI systems or what interventions would most effectively prevent such coordination.

Triggered facilities

The traditional crisis-response mechanisms of central banks rely on discretionary decisions. However, the amplified speed of the next AI crisis will likely overwhelm the existing human-centric stability mechanisms. If private sector AI perceive that the central banks have not adequately engaged with the challenges arising from AI, strategic complementarities may induce them to run — cause a crisis — in cases where they would have stayed — prevented a crisis — if it saw the central bank as well prepared. In other words, the private sector AI's perception of central bank AI preparedness directly affects financial stability.

When considering crisis response mechanisms, the speed of crises implies that a human-in-the-loop approach will not be sufficient, and may even be counterproductive. Therefore, the only viable option may be to to move towards automatic facilities, using either automatic threshold-triggered responses or AI strategic interventions. It seems likely that both will be needed.

Given the speed of AI crises, the authorities should consider moving towards automatic, precommitted liquidity facilities that are activated without human intervention when predefined triggers are reached.

For banks, such facilities might include automatic access to central-bank liquidity (both cash and securities) when certain market indicators reach threshold levels. These facilities might also have to be available to non-banks with large footprints in systemically important markets. They should be designed with safeguards to prevent moral hazard, perhaps including pre-established collateral requirements and haircuts to protect the central-bank balance sheet.

A strategic AI facility implies the central bank already has an AI engine in place, one trained with federated learning by multiple authorities and the AI-to-AI links discussed above. It then would monitor the buildup of vulnerabilities in real time. In times of stress, it would test potential intervention strategies by leveraging the AI-to-AI links to simulate crisis responses, and then implement the optimal intervention, either automatically or with a human on-the-loop.

Such facilities would reduce uncertainty during market stress, prevent destructive AI-driven fire sales, create predictable stabilisation mechanisms and counter the speed advantage of AI systems. By establishing credible, automatic intervention mechanisms, the authorities can influence AI behaviour even before crises emerge, preventing coordination and other destabilising strategies.

Monitoring AI

Many authorities already incorporate AI into their monitoring of the financial system. Our results suggest further avenues for doing so, such as keeping track of which particular AI architectures are used in individual divisions of financial institutions. This includes how they are trained, whether they are internally created, commercial, open source or federated, and what data is used to train them. Regulators should organise this oversight function around the near continuous monitoring of AI use in the financial system, focusing on potential coordination risks.

Monitoring should include a regular assessment of AI deployment across systemically important institutions, the evaluation of common dependencies on AI vendors or data sources, and the identification of emerging synchronisation patterns in AI-driven decision-making. That will allow the authorities to identify the build-up of threats, such as those arising when key operations across banks use similar AI architectures, which suggests a heightened risk of procyclicality.

Particularly important is when the treasury function (liquidity management) extensively uses AI, as is now happening in certain banks. This speeds up the reaction to stress and intensifies crises. Similarly, if the liquidity management operations across banks use similar AI engines, it amplifies the potential for destructive synchronous behaviour, such as increasing the speed and viciousness of crises. If monitoring finds that certain financial institutions can develop their own engines while most cannot, it suggests future concentration and, hence, systemic risk.

Conclusion

Artificial intelligence brings substantial benefits to the financial system: it increases efficiency, lowers the cost of financial intermediation and reduces errors. It also poses new risks to financial stability through malicious use, wrong-way risk, synchronisation and speed.

The likelihood of AI-induced financial crises is inversely correlated with the authorities' understanding and effective deployment of AI technology. If the financial regulators and central banks fail to respond adequately, they risk becoming ineffective in an AI-dominated financial landscape. However, when these authorities opt to harness AI effectively by developing expertise with AI, setting up AI-to-AI links, developing triggered facilities and incorporating AI into their monitoring frameworks, they will become more effective stewards of the financial system.

References

Danielsson, J (2024), "The one-in-a-thousand-day problem", VoxEU, 24 December. https://cepr.org/voxeu/columns/one-thousand-day-problem

Danielsson, J, R Macrae and A Uthemann (2023), "Artificial Intelligence and Systemic Risk", *Journal of Banking and Finance* 140: 106290. https://doi.org/10.1016/j.jbankfin.2021.106290

Danielsson, J and A Uthemann (2024), "On the use of artificial intelligence in financial regulations and the impact on financial stability", SSRN, 5 June. https://doi.org/10.2139/ssrn.4604628

Gambacorta, L and V Shreeti (2025), "The AI Supply Chain", BIS Papers No 154.

International Monetary Fund (2024), *Global Financial Stability Report: Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability* (October), Washington, DC. https://www.imf.org/en/Publications/GFSR/Issues/2024/10/22/global-financial-stability-report-october-2024

Norvig, P and S Russell (2021), *Artificial Intelligence: A Modern Approach*. London: Pearson.

Scheurer J, M Balesni and M Hobbhahn (2024), "Large language models can strategically deceive their users when put under pressure", Technical Report. https://doi.org/10.48550/arXiv.2311.07590

About the author(s)

Jón Daníelsson is a SUERF Fellow, the Director of Systemic Risk Centre and Professor of Finance at the London School of Economics. Since receiving his PhD, Jón's work has focused on how economic policy can lead to prosperity or disaster. He is an authority on both the technical aspects of risk forecasting and the optimal policies that governments and regulators should pursue in this area. Jón has written three highly regarded books: The Illusion of Control (Yale University Press, 2022), which was included on the Financial Times "Best books of 2022" list; Financial Risk Forecasting (Wiley, 2011); and Global Financial Systems: Stability and Risk (Pearson, 2013). He has also contributed numerous academic papers on systemic risk, artificial intelligence, financial risk forecasting, financial regulation and related topics.

Andreas Uthemann is a Principal Researcher in the Financial Markets Department of the Bank of Canada. His main research interests are in financial economics with a particular focus on market structure and design. Before joining the Bank of Canada, Andreas was a postdoctoral researcher at the London School of Economics. He received his PhD in Economics from University College London.

SUERF Policy Notes and Briefs disseminate SUERF Members' economic research, policy-oriented analyses, and views. They analyze relevant developments, address challenges and propose solutions to current monetary, financial and macroeconomic themes. The style is analytical yet non-technical, facilitating interaction and the exchange of ideas between researchers, policy makers and financial practitioners.

SUERF Policy Notes and Briefs are accessible to the public free of charge at https://www.suerf.org/publications/suerf-policy-notes-and-briefs/.

Designed by the Information Management and Services Division of the Oesterreichische Nationalbank (OeNB)

SUERF Secretariat
c/o OeNB, Otto-Wagner-Platz 3A-1090 Vienna, Austria
Phone: +43 1 40 420 7206
E-Mail: suerf@oenb.at
Website: https://www.suerf.org/