AI and financial stability

Jón Daníelsson
London School of Economics

modelsandrisk.org/AI

10 December 2025

Banking Supervision Research Conference 2025

Recent AI cases

a. AI instructed to comply with securities laws and maximise profits — proceeded to do insider trading and lie about it (Scheurer, Balesni, and Hobbhahn 2024)

b. Bank of Canada January 2025: 25% US tariffs impact = GDP ↓ 6%
DeepSeek: 12 seconds to build own model = GDP ↓ 4% (range 1–8%) (Kelton 2025)

c. Google AI solved superbug problem in 2 days that Imperial College worked on for years — provided 5 hypotheses, all valid, one novel

Our notion of AI

- I see it as a *rational maximising agent*
  - Following Russell and Norvig's (2020) taxonomy
- Because that resonates with how economists think
- A computer algorithm performing tasks usually done by humans
- It differs from machine learning and traditional statistics
- Provides quantitative analysis, gives recommendations and makes decisions
- While generative AI takes most of the oxygen, other types can be more important in finance

Recommend Russel (2019) for a general overview of AI

Human-AI interaction levels

- *Human in-the-loop* — AI assists, human makes all decisions
- *Human on-the-loop* — AI operates, human supervises and can intervene
- *Human out-of-the-loop* — AI acts autonomously

- In many domains, the financial sector is largely in-the-loop, with movement toward on-the-loop
- Out-of-the-loop raises serious regulatory and systemic risk concerns

AI Risks and Issues

Main AI societal risks

Weidinger et al. (2022), Bengio et al. (2023), Shevlane et al. (2023), Bostrom (2014), Arthur (1994) and Varian (2018)

1. Trust, entrenchment and over-reliance
2. Misalignment with human values
3. Malicious use and jailbreaking
4. Market power and concentration
5. AI feedback loops and model collapse
6. Shadow AI and ungoverned deployment
7. Operational risks from vendor dependence
8. Bias and discrimination
9. Misinformation and disinformation
10. Autonomy and loss of human control
11. Erosion of privacy

1. Trust, entrenchment and over-reliance

- AI builds up trust by being good at simple tasks that play to its strength
- We may end up with the AI version of the Peter principle
- Becomes essential no matter what senior decision-makers wish
- Usually not credible when someone says
  - "We would never use AI for X"
  - "We will always have a human in the loop"
- So long as it delivers significant cost and efficiency savings

2. Misalignment — Objectives

- How do we ensure AI does what it is supposed to do?
- Impossible for problems that are effectively infinitely complex
- Note analogy with the legal system and incomplete contracts
- And disastrous outcomes
  - EURISKO and the canon of human knowledge and ethics
- Meting objective may depend on intuition to find analogous problems from seemingly unrelated domains that can provide creative solutions
  - AI is not good at intuition

1/2. Trust and misalignment — Explainability

- We demand human decision-makers explain their decisions
- Explainable AI
- Limited but improving explainability in frontier models
- Still insufficient for many regulatory requirements

2. Misalignment — Hallucination

- AI may generate outputs not based on reality when asked to extrapolate to cases not in its training dataset
- Difficult to detect except for experts
- We can reduce hallucination, but do not eliminate it

3. Malicious use — Manipulating AI engines

- Jailbreaking and prompt injection
- Exploiting AI vulnerabilities
  - Adversarial inputs to cause misclassification
  - Data poisoning attacks
  - Model extraction and reverse engineering
- No comprehensive defence yet exists

3. Malicious use — Attacks on society

- Find legal or regulatory loopholes
- Crime and fraud
- Terrorism
- Nation-state attacks on financial infrastructure

3. The defender's dilemma

- Attackers need one vulnerability, defenders must monitor everywhere
- Monitoring gaps between regulatory silos
- Innovation outpaces regulatory frameworks
- Defence requires vastly more resources than attack
- The job of malicious agents becomes easier the better AI becomes

Asymmetry in computing power grows
with system complexity and AI capability

6. Shadow AI and ungoverned deployment

- Employees using ChatGPT, Claude without authorisation
- Uploading proprietary data, getting unvetted advice
- Impossible to prevent entirely
- Creates data leakage and compliance risks
- Requires governance frameworks, not just bans

Limits of kill switches

- Traditional circuit breakers suspend trading
- AI systems cannot simply be turned off
  - Systems depend on AI to function
  - Shutdown might cause more disruption
  - Multiple interacting AIs across institutions
- AI may anticipate and adapt to shutdown attempts

AI Economic/Financial System Risks and Issues

Based on
Daníelsson and Uthemann (2025)

Main AI economic/systemic risk channels

1. Endogenous system responses
2. Objectives (and speed/misalignment)
3. Strategic complementarities and coordination
4. Monoculture and market structure

1. Data

- AI depends on data
- Financial system generates terabytes of data daily
- It is often badly measured and confined to silos
- Little data on most important events as crises are rare (1 in 43 years)
- Observed data lives in the centre of the outcome distribution, not the tails
  - No data on liquidity impact of large trades

AI will not learn much about tail events — crises

1. Unknown-unknowns

- All crises have the same common fundamentals
  - Liquidity, leverage, one day in 1,000 problem, opacity
- While every crisis is unique in detail
- Axiomatic, as else they (hopefully) would have been prevented
  - We don't know about crises successfully prevented
- Crises are *unknown-unknowns* or uncertain (Knight 1921)

AI will not learn much about tail events (crises) since little data on them

1. Endogenous system responses

- Neural networks are a reduced-form model ("super VAR") needing policy experiments in their training data (Sims 1980)
- The reaction functions of market participants and the authorities is *hidden* until we encounter stress
  - Lucas (1976) and Goodhart (1974)

AI finds it hard to learn about how the system behaves in crisis as there is little data to train on

1. Endogenous system responses — Wrong way risk

- Want the highest explainability/least hallucination for most vital problems
- Extreme financial system outcomes unique
- AI knows little about the most important causal relationships
- The reaction functions (public and private sectors) are mostly unknown
- Such a problem is the opposite of what AI is good for
- AI knowledge is negatively correlated with the importance of decisions

When AI is needed the most, it knows the least — wrong way risk

2. Objectives — Principal-agent

- Regulations align private incentives with society
- Principal-agent problem between authority and institution (one-sided)
- Becomes two-sided (institution – regulator – AI)
- Regular ways to incentivise — carrots and sticks — don't work with AI
- It does not care about punishments, being put into jail or losing its bonus
- Scheurer, Balesni, and Hobbhahn (2024) insider trading

We don't know how to incentivise AI

2. Objectives — What to tell it to do?

- How to ensure AI follows the objectives desired by its human operators?
- AI needs clear objectives — more than a human because of hallucination and lack of context
- Micro rulebook known and relatively immutable on operational time scales
- Macro rulebook is not
- Mutability increases along with longer time scales and severity
- Practical macro objectives not known (except at the highest levels of abstraction)
- The worst case is when the objective is unknown ex-ante and cannot be learned, as is the case with the worst financial crises

How to tell AI what to do when we don't know what we want?

2. Objectives — Speed and misalignment

- AI speeds up analysis and decisions
- Decision loops shorten
- Faster private responses synchronise around shared signals and thresholds

How to tell AI what not to do when it acts so fast?

3. Strategic complementarities

- The best course of action for one player increases incentives for others to do the same
- AI train AI — Models train each other
- Private-sector AI might coordinate on undesirable outcomes
  - Run equilibria
  - Market manipulation

AI systems converge on undesirable equilibria
More correlated tail outcomes

3. Procyclicality

- Credit expansion in booms and tightening in downturns
- Feedback via capital and liquidity ratios

AI can amplify financial cycles through correlated decisions and constraints

4. Risk monoculture

Increasing returns to scale of AI technology

- Risk monoculture drives booms and busts
- Humans are more heterogeneous than AI and hence are more stabilising
- Dual oligopolies emerging:
  - AI models: OpenAI, Anthropic, Google, DeepSeek
  - Financial data: Bloomberg, LSEG, S&P

Neither competition nor financial authorities have appreciated the systemic risk from dual oligopolies in AI and data

4. Market structures and competitive advantage

- The largest banks — GSIBs — find it easier to fund their own neural networks and impose AI on staff
- Middle tier banks with traditional staff, legacy systems and technical debt resort to commodity engines
- The smallest neo institutions with nimble technology stacks may find creative uses for commodity engines
- GSIBs ↑, middle tier ↓, neo banks ↑
  - Leads to increased market concentration
  - Middle tier squeezed out
- Automation can hollow out junior risk and operations roles, weakening oversight capacity over time

AI can further entrench GSIBs and hence increase systemic risk

Technology and crises

- One of the earliest applications of the first transatlantic telegram cable in 1858 was the transmission of stock prices
- Apocryphal accounts say Nathan Rothschild used pigeons to get the first news of Napoleon's defeat at Waterloo in 1815 to manipulate the London stock market
- Portfolio insurance and programme trading contributed to the 19 October 1987 crash
- Similar dynamics contributed to the June 2007 stress event and several recent flash crashes

At the onset of every crisis

- Options when faced with a shock — run or stay — stabilise or destabilise
- If shock is not too serious, optimal to absorb and even trade against shock
- If avoiding bankruptcy demands a swift, decisive action, such as selling into a falling market, do exactly that

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

Important to anticipate what the majority will do and do that first

If AI decides to stay

- Will not sell or withdraw liquidity
- Buy
- AI is a stabilising force

If it wants to run — acts as a crisis amplifier

- Speed is of the essence
- The first to react gets the best prices
- The last to act faces bankruptcy
- Sell, call in loans and withdraw funding as quickly as possible
- Makes the crisis worse in a vicious cycle
- Extreme volatility
- AI will significantly speed up and strengthen responses, making crises particularly quick and vicious

Crisis speed from days or weeks to minutes or hours

Options and Expectations

A. Expectations — black-box limits

B. Mandate and capability

C. Implement their own AI engines

D. AI-to-AI links

E. Triggered central bank standing facilities

F. Reporting and dashboarding

A. Why black-box AI is inevitable

- The authorities have a strong desire for explainability
- That means being able to trace decisions to a specific mechanism, human or automated
- Post-hoc (asking AI why it did X) explanations can help, but they are partial and often incorrect and misleading (nobody really knows how AI made a decision)
- Asking why a model made a decision years is likely impossible
- Demanding transparency by opening models undermines competitiveness and increases the attack surface
- Instead of asking for the impossible, redirect efforts

B. Who takes the lead on AI

- The core function of central banks is monetary policy and financial stability
- AI can threaten financial stability
- The financial stability function should lead on AI
- Not data, IT, innovation or some AI hub

Put AI at the core of the financial stability function

C. Authority AI and human capital

- Authorities need to build up expertise in ML and AI
- Run open-source LLMs
- Train financial stability staff in ML and AI
- Actively hire ML/AI experts
- Set baseline ML/AI literacy expectations for new hires and provide training for incumbents
- Provide secure experimentation environments

C. Authority AI system — Why

- Design and benchmark regulations
- Evaluate and benchmark the effectiveness of interventions
- Gain expertise in AI

C. Public-private partnerships — Setup

- Authorities need AI that matches private-sector AI
- Hard to do in-house
- Outsourcing more likely
  - Private firms set up and run engines and AI-to-AI links
  - As is increasingly common in fraud, KYC, AML and credit risk

C. Learning and confidential data

- Coordinate with other authorities on setting up neural networks
  - Share resources
  - Capture common vulnerabilities
  - Capture silo boundary vulnerabilities
- Use federated learning
  - Train models across organisations
  - Share weights but not data
  - Preserve confidentiality while improving models

D. AI-to-AI links

- Authority AI directly talks to AI in private firms and other authorities
- Benefits
  - Benchmarking
  - Test compliance
  - Perhaps how they decide on a loan application
  - Respond to a shock
  - And a crisis intervention
- Might not need data sharing

E. Authority response window

- Recall the last section on crises and their speed
- Faster private reactions compress intervention windows

E. Triggered central bank standing facilities

- Central banks prefer discretionary facilities — *Constructive ambiguity*
- Facilities with predetermined rules for immediate triggered responses
- Can rule out bad equilibria
- Pre-committed, transparent triggers can avoid bad equilibria under rapid, AI-amplified sell-offs; design to minimise gaming
  - If private engines know CBs will intervene if prices drop by X
  - Will not coordinate on strategies that are only profitable if prices drop $> X$
  - Keeps moral hazard low
- If poorly designed, allow gaming and, hence, moral hazard

F Regulation and monitoring today for AI

- Today: PDF reports, database dumps and conversations
- AI can open up a new dimension in the authority-private-sector interaction
  - Regulations more robust and efficient
  - Better monitoring of misconduct and stress
  - Finding latent systemic risk
  - More efficient crisis interventions

F. Reporting and dashboarding

- What's in use: which models are used, where they run, and for what functions
- Changes and issues: major updates and reported incidents
- Signs of dependence and synchronisation: reliance on the same vendors/models, similar liquidity or risk moves across firms

Conclusion

- AI broadly positive for users and supervisors of the financial system
- Raises new, poorly understood systemic risks
- Likelihood of crises negatively correlated with authorities' understanding and use of AI

Monetary policy, financial stability, supervision — should be in charge of AI, not support divisions like IT, data, library, innovation

If the authorities do not effectively engage with AI, crises become more likely