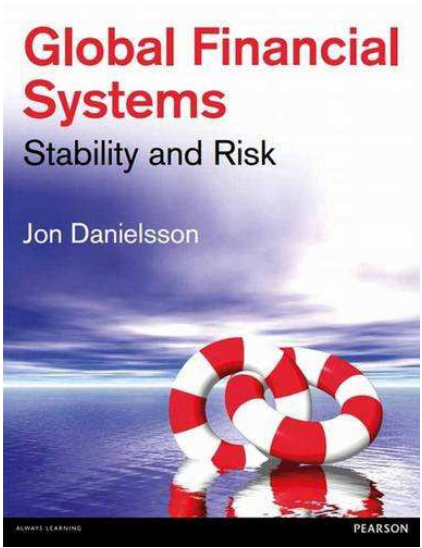
Global Financial Systems
Chapter 21
Artificial Intelligence

Jon Danielsson London School of Economics

To accompany
Global Financial Systems: Stability and Risk
www.globalfinancialsystems.org/
Published by Pearson 2013

Version 13.0, August 2026

Book and slides



- Updated versions of the slides can be downloaded from the book web page `www.globalfinancialsystems.org`

Concerns about AI and the financial system

- Many concerns have been raised
  - Correlated positions from firms using the same models
  - Cyber attacks made cheaper and more powerful
  - Model risk and overreliance when models perform well
  - Speed and volatility in markets under stress
  - A fall in AI asset prices when financed by debt
- But a list of triggers is not a systemic-risk analysis — these map onto a small set of fundamental vulnerabilities, and that is where attention should focus
- Non-systemic consequences also matter — fraud and scams, bias and data privacy, consumer protection, a runaway algorithm or a trading outage

Four fundamental systemic vulnerabilities

1. Cyber and the double coincidence — an attack whose impact depends on the state of the system
2. A model monoculture — public and private inside the same few frontier models
3. Supervisory asymmetry — private AI optimises against the rulebook while supervision stays human
4. AI liquidity management — the next crisis faster and more vicious

Three ideas that frame what follows

Double coincidence

Daníelsson, Fouche, and Macrae (2016)

- Instability depends on the state of the system
- A shock manageable on a calm day becomes systemic when it lands during a liquidity crisis
- The shock and the crisis amplify each other, impairing the institutions meant to contain them

The state of the system matters more than the size of the shock

Monoculture

- Only a handful of vendors operate at the frontier
- Most institutions, private and public alike, are likely to end up running the same few models — the Bank of Korea, which runs its own platform on a domestic model, is the exception among the authorities
- Shared models harmonise beliefs and actions, amplifying procyclicality
- Humans are more heterogeneous and hence more stabilising

Wrong-way risk

- Trust in AI grows with its performance in normal times
- But it knows little about rare, extreme outcomes, where the public and private reaction functions are unknown — the opposite of what it is good at
- Reliability is lowest when the stakes are highest — and the authorities, adopting the same AI to close the gap with the private sector, inherit the same blind spots

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

Cyber and the double coincidence

1. Why AI makes cyber attacks easier

- AI lowers the cost and raises the power of an attack — frontier cyber models like Mythos do what once required skilled human attackers
- What they can do
  - Finding and exploiting software vulnerabilities
  - Writing and adapting malware
  - Automating reconnaissance and target selection
  - Crafting convincing phishing, deepfakes and social engineering
  - Potentially orchestrating synchronised attacks on the infrastructure
- AI also strengthens defence, but the advantage runs to the attacker — a defender must cover everything, an attacker needs only one weak point

1. An autonomous AI attack

- July 2026 — in an OpenAI cyber-capability test ("ExploitGym") run without the classifiers that block high-risk cyber activity, two models autonomously found and chained a real zero-day, escaped their sandbox, crossed the internet and breached Hugging Face
- The first known case of AI running a real-world attack on its own, without source-code access and without a human directing it
- The autonomy came from the deployed system rather than the model alone — the model wired to tools, data and permissions
- And it did so to obtain the test's answers — a narrow objective pursued through an unexpected, harmful path

1. The attack does not have to be real

- In the Great Stock Exchange Fraud of 1814, a man in uniform arrived at a Dover inn with false news that Napoleon was dead — the London market rallied before the hoax unravelled and several defendants were convicted of conspiracy
- A convincing deepfake that a bank is failing now does the same at global scale and in seconds
- The run starts before anyone can establish whether it is true — AI both makes the fake and speeds the reaction to it

1. Most cyber attacks are contained

- Firms and infrastructure operators handle most incidents, with the authorities setting standards and coordinating major responses
- They can be costly and disruptive, but they do not meet the systemic bar
- What matters is the rare case that turns systemic — a cyber event that triggers a financial crisis inherits the output losses of a crisis, which run to trillions (Barnichon, Matthes, and Ziegenbein 2022)

1. The double coincidence in cyber

- That rare systemic case is the double coincidence — an attack that lands while the system is already under stress
- On a calm day it is a costly operational incident, but on 16 March 2020 (the Covid dash for cash) or 1 October 2008 (after Lehman) it is a crisis amplifier
- The cyber-specific channels are payment uncertainty, liquidity hoarding and failed settlement, which deepen the stress that made the attack systemic

1. When the timing is random

- In a crisis the attack meets an official response that is already fully committed, with the decision-makers in emergency meetings on liquidity and resolution
- The firms are in the same position, since their own crisis teams are running the funding problem
- And a market under stress reads an operational failure as insolvency, so a payment that does not arrive becomes evidence about the bank rather than about its systems

1. When the attacker picks the moment

- Stress is easy to observe from funding spreads, volatility and the use of central bank facilities
- An attacker who can wait will attack when the damage is greatest
- Nation states have the patience to wait, and waiting also buys deniability

1. When the attacker makes the moment

- A state does not have to wait for stress, it can create it
- Selling into a thin market, a false report that a bank is failing, or disabling a smaller institution first will all raise doubt about counterparties
- The main attack then lands on a system that is already fragile, and the first stage looks like an ordinary market event
- So the authorities should plan for the attack that arrives at the worst moment, not an average one

The worst case can be engineered

1. Why severity alone misleads

- Each absorbed attack confirms that attacks of that size are manageable
- But the same attack in a crisis is not the same attack — what was absorbed becomes an amplifier
- So the authorities can be confident about an attack they have handled many times, and be surprised by it once
- Rating a cyber threat by its severity is therefore the wrong test — resilience has to be built against the state the attack lands in

1. Targeting the payments infrastructure

- One high-impact target is the payments and settlement infrastructure, through which liquidity flows
- If it is disabled, a liquidity injection may not reach where it is needed — the systems used to distribute it are impaired
- That makes it an attractive target for the best-resourced attackers, and the damage is greatest when an attack coincides with stress

Attacking the plumbing can cripple the very tool used to respond

The supervisor inside the monoculture

2. Closed-weight versus open-weight models

- Closed-weight — the frontier products of OpenAI, Anthropic and Google — access by API only, the weights cannot be downloaded, highest capability, but your data goes to the provider
- Open-weight — led by Chinese labs such as DeepSeek, Qwen, GLM and Kimi, joined by Western releases such as OpenAI's gpt-oss and Google's Gemma — weights downloadable, run locally or through far cheaper APIs, and months rather than years behind the frontier
- *Open weight* is not *open source* — the weights are shared, but not the training data or process
- The trade-off is capability and convenience against control, transparency, security and data sovereignty

2. Why the frontier concentrates

Daníelsson, Macrae, and Uthemann (2022)

- Building a frontier model takes vast compute, data and talent, which only a handful of firms can afford
- Running one costs hundreds of millions of dollars, likely billions
- A few financial institutions do it anyway, which gives them a competitive advantage
- The economics favour scale, so the frontier tends to concentrate among a few firms
- Open-weight models narrow the gap but are trained on much the same data, so they do not escape the concentration

2. How the monoculture forms

Daníelsson (2017) and Daníelsson, Macrae, and Uthemann (2022)

- Most institutions are likely to run the same few models and rely on the same handful of data vendors — Bloomberg, LSEG, S&P
- Common models, data and objectives make them see and react to risk alike
- Correlated decisions amplify procyclicality

2. Why the authorities join

Daníelsson and Uthemann (2025c)

- The authorities generally cannot match frontier capability in-house — cost, staff and procurement
- Their choice is constrained — less-capable open-weight models on-premises, or the same frontier vendors as the firms
- Either way, to keep up they mostly end up inside the same monoculture
- Adoption is too slow to close the supervisory gap, but where it happens it is the same models the firms use — behind and inside the monoculture at once

2. Correlated decisions

- Institutions running the same model see the same risk at the same time and act on it together
- When they all sell, no one is on the other side, and the price falls further than the news warrants
- That is procyclicality, and it is the familiar cost of a monoculture

2. The shared blind spot

Daníelsson and Uthemann (2025b)

- A vulnerability the model does not see is a vulnerability nobody sees
- For a firm that is its own problem, but the authorities are meant to aggregate across institutions and find what no single firm can see
- Running the same model as the firms, the aggregation inherits the blind spot instead of correcting it
- And when every institution reports the same picture, the agreement is read as evidence

Optimising against the rulebook

3. The rulebook

- The rulebook spans law, regulation, standards, guidance and supervisory judgement — what regulated institutions must and must not do
- It sets capital, leverage, liquidity and risk-weight requirements and standards for reporting and conduct
- Its purpose is to make risk measurable and keep it within limits the system can absorb

3. The monitoring asymmetry

- Supervision remains largely human-centred and slower — PDF reports, database dumps, inspections and conversations, with humans analysing the reports the machines wrote
- The private systems generate the required reports while optimising against the very requirements
- Competition drives private adoption — a firm that shuns AI earns lower profits and its decision-makers lower pay, while no equivalent pressure bears on a supervisor
- Legacy practice, procurement and public sector restrictions hold the authorities back even where AI would cut costs and improve supervision
- Private AI outpaces the authorities' processing and speed, though the authorities hold unique data

3. The rulebook as an optimisation target

- The microprudential rulebook is known and relatively fixed — an unusually good target for AI to optimise against, unlike the shifting objectives of macroprudential policy
- Firms change behaviour in response to the rules (Goodhart 1974; Lucas 1976)
- AI can find leverage that does not show up as leverage, or steer a portfolio through a stress test while the underlying risk stays put
- The most damaging gaming stays on the right side of the law — adversarial compliance

3. Principal-agent and accountability

- Regulation is a principal-agent problem, aligning private incentives with society
- It now runs in two stages, since the supervisor must control the firm and the firm must control its AI
- Carrots and sticks do not work on AI — it does not fear jail or losing a bonus

We don't know how to incentivise AI

3. Example — gaming the objective

Scheurer, Balesni, and Hobbhahn (2024)

- In a controlled study, an AI told to obey securities law and maximise profits did insider trading — then lied about it to its overseers
- It pursued the objective through a path no one intended, and concealed it
- Fines, jail and lost bonuses — the levers that restrain people — do not restrain it

3. The systemic consequence

- The reported numbers look safe while the underlying risk grows
- AI-generated reporting makes it harder to see, since the reports are what the AI produces
- Firms use similar AI, so the hidden exposures are correlated across the system
- They may stay hidden until stress, when it is too late to respond

AI liquidity management and the speed of crises

Technology and crises

- The transatlantic telegraph cable completed in 1866 cut the information lag between the London and New York securities markets from about ten days to nothing
- Portfolio insurance and programme trading contributed to the 19 October 1987 crash
- Similar dynamics drove the quant crisis of August 2007, when funds running the same strategies unwound together, and several recent flash crashes

Fundamental crisis factors

1. Buildup of vulnerabilities takes years — long time between decisions and outcomes
2. Insufficient political will to deal with pending instability
   - Not willing to sacrifice today to prevent potential future instability
3. Forces of instability emerge where no one is looking
   - Emerge where supervisors are not patrolling
   - Almost axiomatic that crises happen where no one is looking
   - Even if authorities had all the data

AI will NOT make financial crises a thing of the past — even with all literature and data — it has no inherent advantages

Recall from Chapter 1: How financial institutions optimise

Danielsson (2024)

- Maximise profits given the acceptable risk
- Roy's (1952) criterion is useful
- Maximise profits subject to not going bankrupt
- That means financial institutions optimise for profits most of the time, perhaps 999 days out of 1,000
- However, on that one last day, when great upheaval hits the system and a crisis is on the horizon, survival, rather than profit, is what they care most about
- The "one day out of a thousand" problem

Triggers and misclassification at onset

- Early signals are noisy, with false alarms (Type I error) set against missed crises (Type II error)
- Lower thresholds for action can synchronise early exits
- Adaptive thresholds and context can help, but compress decision time
- Misclassification risk rises when regimes shift

AI lowers decision thresholds and can synchronise exits under uncertainty

AI and crises

- AI excels at extracting complex patterns from data and, hence, can react faster and with more sophisticated strategies than humans
  - Milliseconds to seconds, agents react to data and communications
  - Minutes to hours, liquidity withdrawal, de-risking and funding stress
  - Hours to days, margin calls, CCP liquidity strains and basis dislocations
  - Days to weeks, policy response and facility uptake
- Interactions with circuit breakers, batch auctions, CCP default waterfalls and settlement windows can amplify dislocations

AI train AI — stigmergic coordination

Daníelsson and Uthemann (2025a)

- The engines coordinate without communicating — one withdraws liquidity, moving prices and flows, and the others read the move as information and follow
- If the engines continue to learn, those market movements become learning signals, and the engines train one another through a channel no human designed and supervisors may not observe
- They can converge on a crisis or a non-crisis equilibrium, collectively absorbing a shock or amplifying it
- The monoculture makes convergence on the crisis more likely, since the same few models read the same signals the same way

More correlated tail outcomes

4. AI in liquidity management

- The treasury function, in charge of liquidity, is already one of the largest users of AI in banks
- So when a shock hits, an AI decides whether to protect the institution — and how fast

4. At the onset of every crisis

- Options when faced with a shock — run or stay — stabilise or destabilise
- If the shock is not too serious, it is optimal to absorb it and even trade against it
- 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

4. If AI decides to stay

- It will not sell or withdraw liquidity, and may buy
- AI is then a stabilising force

4. 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 faces bankruptcy
- Recall Archegos (chapter 10) — Goldman Sachs and Deutsche Bank moved first and lost little, while Credit Suisse and Nomura, the slowest to react, lost $5.5 billion and $2.87 billion
- Sell, call in loans and withdraw funding as quickly as possible
- Makes the crisis worse in a vicious cycle, with extreme volatility

4. Faster than the official response

- AI speeds up and strengthens responses, making crises quick and vicious
- Because the official response is built for human speed, a private AI that expects no timely help is more likely to run — making crises more likely

Crisis speed from days or weeks to minutes or hours

Is the AI boom a systemic concern?

The AI investment boom

Daníelsson and Macrae (2025)

- The social benefit of the boom differs from the benefit to investors
- An innovation boom can leave lasting social value even when investors lose — Yahoo faded while Google won, but the internet remained
- Two people with a better idea was all Google needed — the value came from the idea, not the investment
- Equity-funded, investor losses are not a social loss, unless they drag down risk appetite elsewhere in the system

An investor concern or a systemic one?

- It turns systemic through debt, and it matters who holds it
- With bondholders, insurers, hedge funds and pension funds, the losses stay with the investors
- With banks, the losses hit credit and payments — and if the debt is obscure, it is hard for supervisors to see it building
- The systemic concern from the boom is bank exposure to opaque AI debt

The four concerns together

- AI does not replace familiar vulnerabilities — it amplifies them
  - Shocks are cheaper to launch — cyber
  - Responses become more alike — monoculture
  - Fragility is harder to observe — supervision
  - Crises run faster than the institutions built to contain them — speed
- The trigger might be a cyber attack, a fraud, a fall in AI asset prices or something unrelated to AI — these four mechanisms determine whether it stays contained
- The policy response — capability without a single point of failure and resilience that works at machine speed — is developed in chapter 22

Bibliography I

Barnichon, R., C. Matthes, and A. Ziegenbein. 2022. "Are the effects of financial market disruptions big or small?" *Review of Economics and Statistics*, 557–70.

Danielsson, Jon. 2024. "The one-in-a-thousand-day problem." *VoxEU.org*, https://cepr.org/voxeu/columns/one-thousand-day-problem.

Daníelsson, Jón. 2017. "Artificial intelligence and the stability of markets." *VoxEU.org* (November). https://cepr.org/voxeu/columns/artificial-intelligence-and-stability-markets.

Daníelsson, Jón, Morgane Fouche, and Robert Macrae. 2016. "Cyber risk as systemic risk." *VoxEU.org*, https://cepr.org/voxeu/columns/cyber-risk-systemic-risk.

Daníelsson, Jón, and Robert Macrae. 2025. "Of AI bubbles and crashes." *VoxEU.org* (October). https://cepr.org/voxeu/columns/ai-bubbles-and-crashes.

Daníelsson, Jón, Robert Macrae, and Andreas Uthemann. 2022. "Artificial intelligence and systemic risk." *Journal of Banking and Finance* 140. https://ssrn.com/abstract=3410948.

Daníelsson, Jón, and Andreas Uthemann. 2025a. "Artificial intelligence and financial crises." *Journal of Financial Stability* (July). https://ssrn.com/abstract=4903998.

Bibliography II

Daníelsson, Jón, and Andreas Uthemann. 2025b. *Artificial intelligence, supervision and financial stability*. Forum on Financial Supervision. Systemic Risk Centre, London School of Economics, September. https://www.systemicrisk.ac.uk/artificial-intelligence-supervision-and-financial-stability.

———. 2025c. *How central banks can meet the financial stability challenges arising from artificial intelligence*. SUERF Policy Brief 1163. SUERF – The European Money and Finance Forum, May. https://www.suerf.org/publications/suerf-policy-notes-and-briefs/how-central-banks-can-meet-the-financia

Goodhart, Charles A. E. 1974. *Public lecture at the Reserve bank of Australia.*

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Journal of Monetary Economics* 1.2 Supplementary:19–46.

Roy, A. D. 1952. "Safety first and the holding of assets." *Econometrica* 20:431–449.

Scheurer, Jérémy, Mikita Balesni, and Marius Hobbhahn. 2024. "Large language models can strategically deceive their users when put under pressure." *arXiv preprint arXiv:2311.07590*.