The London **Foundation** for Banking & Finance | CSFI Centre for the Study of Financial Innovation

Institute and Faculty of Actuaries

A charity incorporated by Royal Charter

It's still not magic:

Framing the risks facing financial services in the Gen AI era

by Keyur Patel

About The London Foundation for Banking & Finance (LFBF)

The London Foundation for Banking & Finance (LFBF) is a registered charity incorporated by Royal Charter. It was founded in 1879 by a group of City-based bank workers who came together to establish leadership and professional practice principles for the industry. Today, its charitable purpose is the advancement of knowledge of and education in financial services, and the conduct and publication of research for public benefit. It awards Chartered status to individuals who demonstrate the highest level of professional competence and is also home to the Centre for the Study of Financial Innovation (CSFI): a think-tank focused on the challenges and opportunities facing the finance sector.

Our heritage

The London Foundation for Banking & Finance started life back in March 1879, when a group of bank workers came together to establish leadership and professional practice principles for the industry. They created the first Institute of Bankers in England and Wales to offer educational resources to anyone working in the sector.

Over the years, the organisation developed its own industry-leading qualifications to create a gold standard of banking and financial education. It also established itself as a leading voice in the banking world, providing invaluable insights into all areas of the industry and promoting the highest standards of professional competency. Today, LFBF exists to support the advancement of knowledge and education in financial services.

Previously The London Institute of Banking and Finance (LIBF), the charity was renamed 'The London Foundation for Banking & Finance (LFBF)' following the acquisition of LIBF's education and training activities in March 2023 by IU Group. LIBF's education and training activities now continue under a new wholly owned UK subsidiary of IU Group – LIBF Limited – which changed its name to Walbrook Institute London Ltd.

Contents

Preface

The London **Foundation** for Banking & Finance | CSFI Centre for the Study of Financial Innovation

Opportunity and risk often go together. But as AI reaches across every part of the financial services ecosystem – and indeed, every part of society – they are especially hard to separate. This report introduces a new framework for understanding AI risks; one which, at its core, is about the trade-offs that give rise to "uncomfortable tensions".

The CSFI – which is now part of The London Foundation for Banking & Finance – first looked at this question in 2019, when attention was mainly on classical machine learning and deep learning. Since then, generative AI and large language models have transformed the landscape. They have lowered barriers to use, make AI feel relatable and trustworthy, and are increasingly embedded in how financial institutions think and work. As the report discusses, this matters because AI outputs can be useful, confident, and wrong at the same time. In finance, 'mostly right' may be good enough for some tasks, and dangerous for others.

The framework in this report identifies nine risks which are grouped into three broad categories: 'outcomes', 'operating environment' and 'system'. We believe this is a useful way to trace how AI risk moves through the financial services ecosystem: from the outcomes experienced by customers and society, to the environment in which firms deploy AI, to the system-level dynamics through which risks can scale and spread.

It also draws out the uncomfortable tensions – for example, that the same machinery can widen inclusion and sharpen exclusion; that 'human in the loop' is not the same as human control; and that concentration is baked into how AI systems are built. Many of the thorniest risks in this report are ecosystem risks. These are not only those explicitly labelled 'system' level, but any risk that can scale, spread, or reinforce other risks across the financial system. They often involve choices that look sensible for individual firms, but can create shared dependencies and fragility across the system.

In 2024, the CSFI became part of The London Foundation for Banking & Finance, a registered charity incorporated by Royal Charter with a purpose to advance knowledge and education in financial services for the public benefit. Understanding the impact of AI on financial services is of the highest importance – and our report is very timely, amid recent scrutiny of the landscape (including, notably, by the House of Commons Treasury Committee). We are most grateful to the Institute and Faculty of Actuaries for all their support in this endeavour.

Keyur Patel
Report author, Research Associate,
The London Foundation
for Banking & Finance

Harry Weber-Brown
Director of Partnerships
The London Foundation
for Banking & Finance

Foreword

The IFoA is delighted to partner with the LFBF on one of the great challenges of our time, not only for financial services but to all society: AI and related emerging technologies.

The report "It's still not magic: framing the risks facing financial services in the Gen AI era" covers all of the key issues. However, its power comes from the fact that it draws widely on the testimony of market participants who are grappling with these issues in real time.

A key feature of the report is that it brings out the duality of opportunity and threat, setting out five tensions which demonstrate the ethical, economic and technical utility of AI. The framework set out in this report argues that AI risks need to be addressed as trade-offs. As it states, "the question is not whether these risks can be eliminated, but how much risk we are willing to live with in exchange for the benefits".

Risk management is at the heart of what the actuarial profession does. The advent of generative AI and its successor technologies represents a major challenge. This challenge is not just technical, but also ethical, and is one that professions such as ours must embrace. The pace of change in this area means that regulation will inevitably struggle to keep pace. As such, it is important that professional judgement plays a pivotal role in managing the risks and trade-offs that the report so starkly sets out.

AI is a defining force of our time, and one which is at the heart of the IFoA's strategy. The IFoA's Artificial Intelligence and Emerging Technologies Practice Board is exploring how transformative technologies are reshaping actuarial practice and influencing broader societal systems. It is also exploring what we need to do to seize the opportunities and manage the risks associated with AI adoption. With our unique combination of technical skill, communication and professional oversight, actuaries must play a key role making sure that AI is working as it should.

We commend this report to the wider financial services community as a snapshot of the state of AI, and in setting out a framework to assess the complex trade-offs we will need to make as these technologies play an increasing role in economy and society.

Paul Sweeting FIA C.Act,
President,
Institute and Faculty of Actuaries (IFoA)

Executive summary

This report introduces a new framework for making sense of AI-related risks in financial services, in a landscape transformed by the widespread adoption of generative AI. It builds on a framework first developed in a 2019 CSFI report, reproduced in the appendix. The risks are grouped under three broader headings:

- **'Outcomes'** captures risks in how AI affects customers and society – including fairness, trust, and the reliability of outputs.
- **'Operating Environment'** captures risks in the conditions under which firms use AI – including cyber threats, human capability gaps, and pressure to deploy too quickly.
- **'System'** captures risks that arise when AI reshapes rules, market incentives, and shared dependency across the financial system.

These categories are not rigid: many AI risks cut across customers, firms and markets. But they provide a useful way to trace how AI risk moves through the financial services ecosystem – from the outcomes experienced by customers and society, to the operating environment in which firms deploy AI, to the system-level dynamics through which risks can scale and spread.

The framework is presented on the next page. In short, it splits AI risks into nine buckets:
1) Fairness; 2) Trust; 3) Truthfulness; 4) Security;
5) Understanding; 6) Overreliance; 7) Governance;
8) Concentration; 9) Contagion.

Three wider risks are shown separately: 'Workforce Disruption', 'Environmental Impact', and 'AI Asset Bubble Risk'. These sit outside the core framework, as broader AI-related risks that the sector may experience, finance or amplify. Together, they point to a wider AI economic transition that financial services will not just use, but help fund, insure and absorb.

The framework treats AI risks as trade-offs, not standalone downsides. Often, the risk is the shadow side of the benefit: sharper prediction, deeper personalisation, greater scale, greater complexity, and more autonomous decision-making. The question is not, then, whether these risks can be eliminated, but how much risk we are willing to live with in exchange for the benefits.

As part of this research, the CSFI conducted a survey of 78 senior financial services practitioners and observers.

- **70%** of respondents agreed that *"Risks arising from the use of AI are among the greatest risks facing my sector over the next five years"*, compared to 17% who disagreed.
- **75%** of responds agreed that *"The risks posed by AI to my sector have increased substantially since generative AI technologies have become widely available"*, compared to 10% who disagreed.
- The top three risks highlighted by respondents to the survey were *1) Cyber Threats and Privacy Breaches; 2) Misleading Outputs and Hallucinations;* and *3) Talent or Knowledge Gaps.*

Five uncomfortable AI tensions

Throughout the analysis in this report, five 'uncomfortable tensions' repeatedly arise. They are not isolated risks, but deeper conflicts that cut across them – where the same features that make the technology valuable also make it difficult to govern, explain, trust, or contain.

1 The same machinery widens inclusion and sharpens exclusion

AI can identify people who were previously invisible to the financial system. But precisely the same ability to detect patterns in richer data can also identify who is less profitable, more expensive to serve, or easiest to leave behind.

2 The useful version is often the socially uncomfortable version

AI systems are most powerful when they are complex, data-intensive or semi-autonomous. Making them simpler, more explainable, less intrusive, or more controllable can strip away the very qualities that made them useful – and, in some cases, break the use case.

3 Human in the loop is not the same as human control

'Human in the loop' is often treated as a safety valve. But human intervention only works if people have the time, expertise, authority, and confidence to challenge the system. Otherwise, it becomes a tick-box exercise around machine-led control.

4 Regulation is reactive because it has to be

Regulators have to anticipate AI risks before they crystallise. But AI is not a fixed technology waiting to be regulated: it evolves through deployment, embeds inside other systems, and changes how decisions are made. By the time risks are visible, the market may already have reorganised around it.

5 Concentration is baked into how AI is built – and vulnerabilities scale with it

AI complicates the assumption that competition naturally disperses risk. It depends on scarce compute, data, talent, and infrastructure. That creates rich targets for attackers and shared exposures that can make separate firms fail in similar ways at the same time.

Gen AI gives these tensions a new force. It lowers the barrier to use, makes AI feel relatable and trustworthy, and creates outputs that may be useful, confident, and wrong at the same time. In finance, that matters because 'mostly right' may be good enough for some tasks – and dangerous for others.

The AI risk map for financial services

Fairness	Trust	Truthfulness	Security	Understanding	Overreliance	Governance	Concentration	Contagion

Outcomes: how AI affects customers and society

Bias and Social Harm	Erosion of Transparency and Trust	Misleading Outputs and Hallucinations
The core trade-off: **Inclusion through sharper prediction, versus exclusion through sharper discrimination.** *The danger is optimisation working too well.*	The core trade-off: **Performance through complexity, versus trust lost if decisions cannot be explained.** *A technical explanation is not the same as a human explanation.*	The core trade-off: **Useful answers in messy situations, versus confident answers that may be false.** *The system may be persuasive when it is wrong.*

Operating environment: how AI changes the conditions firms operate in

Cyber threats and privacy breaches	Human talent and knowledge gaps	Rushed or poorly managed adoption
The core trade-off: **More data-driven defence, versus more data-rich targets.** *The tools that improve security can also enlarge the prize for attackers.*	The core trade-off: **Powerful tools for more people, versus fewer people able to challenge them.** *"Human in the loop" only works if the human understands the loop.*	The core trade-off: **Competitive advantage through speed, versus avoidable harm through haste.** *Fear of falling behind can push firms ahead of their controls.*

System: how AI risks scale and propagate

Regulatory gaps and fragmentation	Market distortion and skewed incentives	Shared dependencies and contagion
The core trade-off: **Innovation through flexible rules, versus gaps where old frameworks do not fit.** *Regulation may be supervising yesterday's architecture.*	The core trade-off: **Compounding advantage through scale, versus markets rewarding concentration.** *The front end may look competitive while the back end concentrates.*	The core trade-off: **The compelling logic of common tools, versus the fragility of common failure points.** *What is prudent for each firm can be dangerous for the system.*

Beyond the core framework: wider AI risks touching financial services

Workforce disruption	Environmental impact	AI Asset bubble risk
Job displacement and disruption to skills and career pathways.	AI compute demands raise energy and infrastructure concerns.	Inflated valuations from AI hype.

The AI economic transition

The financial services sector will not just use AI – it will need to fund, insure and absorb its shocks.

Why focus on risks, and why now?

Artificial intelligence is overwhelming in a way that no previous technological breakthrough has been before.

It is overwhelming because of the startling pace of change: capabilities that barely existed a few years ago (fluent large language models and generative AI) are ubiquitous today, for better or worse, and underpin the valuations of many of the world's most valuable companies. It is overwhelming because the societal implications are so difficult to anticipate, and to bound – from the future of work and privacy to the environment and healthcare, and on to possibilities that remain speculative but are no longer fanciful. It is overwhelming because we do not know what the technology will be capable of in five years, or indeed, in six months.

It is already touching every part of the financial services ecosystem – consumers, institutions, and the system as whole – across all its sectors and layers of activity. Few sectors are more obviously primed for AI. At their core, AI applications are usefully understood as prediction machines, and financial services is an industry built around prediction: who will repay, what risks should be insured, which transactions look suspicious, and how markets may move.

All that is true. Yet at the same time, it is not churlish to worry about hype and inflated expectations. This report is a sequel to a 2019 CSFI report, by Keyur Patel and Marshall Lincoln, on risks facing financial services with the growing use of AI (meaning, for practical purposes, machine learning and its subset, deep learning). We called that report "It's not magic", because, beneath the excitement, the mechanics were not mystical, even if they could be opaque. AI was – and still is – rooted in advanced pattern detection.

Gen AI is not magic either. Nor is agentic AI. But whereas the earlier wave of machine learning could still feel magical to the relatively small proportion of people who encountered it directly, Gen AI can feel magical to anyone with an internet connection. Agentic AI takes that one step further: embedding the same magic-feeling interface into systems that can *act*.

There is also the uncomfortable reality that our understanding of how data and computing power becomes the kind of fluent text – and, increasingly, images and video – we can now produce at a touch of a button is hazy at best, even to the world's leading experts. There is a persuasive case that it is more discovery than invention. Whatever one thinks of the claim that the technology is 'just' unthinking next-token prediction, there is no doubt that it is useful.

Risks are really trade offs

The focus of this report is risks – but that is not to play down the opportunities AI can bring to financial services. Much insightful and well-evidenced research has been conducted into these opportunities. The real point to make here is that risks are not standalone downsides – they are trade-offs. They often exist because of the benefits, not in spite of them. In Silicon Valley parlance, they are often features of the technology, not bugs.

One of the main aims of this report is to present a risk framework for understanding how AI affects financial services. It builds on the framework introduced in the 2019 report, which remains relevant, but needed updating for the Gen AI era. The new framework explicitly lays out the trade-offs at the heart of each risk. While many of the risks can be ameliorated through better risk management, governance, and regulation, it is worth making the point explicitly that they cannot be eliminated altogether. In other words, if we want the technology, we may have to live with some of the risk. And so, for each risk, this report looks at an 'uncomfortable tension' – exploring how the downside is, to some degree, baked into the opportunity itself.

'Risks' can mean many different things from different perspectives. The LFBF exists to advance financial knowledge for the benefit of the public, and these are the risks this report focuses on. For example, AI can pose risks to a financial institution because, in not adopting the technology effectively, it could miss out on the opportunity to improve market share. That is real, but it not it is not the kind of risk this framework is primarily concerned with.

What matters here is when that firm-level pressure creates wider consequences: rushed adoption, weak safeguards, harmful customer outcomes, distorted competition, or greater fragility in the financial system. In other words, this report focuses less on the competitiveness of one firm against another, and more on the health of the financial services ecosystem as a whole.

How this report was written

The timing of this report is underlined by the House of Commons Treasury Committee's 2026 report on artificial intelligence in financial services. The Treasury Committee said, bluntly:

"The Financial Conduct Authority, the Bank of England and HM Treasury are not doing enough to manage the risks presented by AI. By taking a wait-and-see approach to AI in financial services, the three authorities are exposing consumers and the financial system to potentially serious harm."

There are different views on that assessment, and this report does not seek to adjudicate between them. It is not a response to the Committee's work, but it is written in the same policy moment – one in which AI risk in financial services has moved from a theoretical concern to a live question of governance, supervision and public policy.

The content of this report is based on the author's ongoing research into AI risks in financial services, including:

- An online survey of 78 senior financial services practitioners and observers.
- Dozens of interviews conducted by the author with subject matter experts.
- Analysis of a wide body of research reports into AI by financial authorities, international organisations, academics, and others.
- The substantive body of evidence that informed the Treasure Committee's report, including 84 written submissions.
- Insights that emerged from the CSFI's 2025 Insurance Banana Skins survey, which showed AI to be the most pervasive risk facing the industry.

Risks are still, ultimately, about humans

The 2019 predecessor to this report concluded that AI risks in financial services are as much about how humans use and interpret the technologies as the technologies themselves. In the Gen AI era, that is truer than ever.

When Joseph Weizenbaum created the rudimentary chatbot ELIZA in the 1960s, what amazed him was not the sophistication of the program (which was crude) but how readily users projected intelligence and understanding onto it. When world chess champion Garry Kasparov was defeated by IBM's Deep Blue in 1997, one of the stories later told about the match was that a strange move by the computer – reportedly the result of a bug[1] - unsettled Kasparov because he read into it a depth of strategy that was not there.

Gen AI simulates intelligence far more convincingly. It can feel like a hugely knowledgeable friend who is eager to tell you what you want to hear – and prone, occasionally and without warning, to making things up altogether. The older risk of humans being seduced by technology becomes dramatically greater when the technology feels as if it has already raced past the Turing test. And this is material in a world where fears of an AI asset bubble are growing, and where speculative memos about AI's impact can move markets[2].

One survey respondent said the main risk AI poses to financial services is: "Paradoxically a simultaneous over-estimation and under-estimation of what the current version of AI can actually do, leading to expensive mistakes and missed opportunities". That feels right. Another said: "My perspective is treating Gen AI the same way as a fresh graduate – trust but verify". Yet what happens when the "fresh graduate" is no longer just assisting the work, but becoming part of how the institution thinks?

This analysis is an independent piece of work, and does not necessarily represent all the views of the Institute and Faculty of Actuaries (who have kindly sponsored it). Any mistakes or interpretative errors are the author's and should not be attributed to the IFoA or the CSFI. I would like to thank everybody who generously contributed their time and expertise.

Became prominent from:	Late 20th century	2010s	Mid-to-late 2010s	2022	2024
	Rule Based Automation (Not AI)	'Classical' machine learning	'Deep' learning	Gen AI/LLMs	Agentic AI
Fundamental new capability	Rules executed at scale. Humans define the logic; machines apply it quickly and consistently	Rules learned from data. System infers patterns rather than following only human-written instructions.	Patterns found beyond human sight. Neural networks detect signals in messy, high-dimensional data.	AI generates human-like content. Produces fluent text, code, images, and video through ordinary language prompts.	AI can act across workflows. Can plan, call tools, update systems and pursue goals over multiple steps.
Fundamental new risk	Bad rules scale. If the logic is wrong, errors are repeated quickly and consistently.	Control over the mechanism weakens. Humans set the goal, but the system learns the route from data.	Power comes with a black box. The model may work, but humans may not know why.	Plausibility can masquerade as reliability. Fluent outputs can persuade, mislead, or fabricate.	Delegation can outrun control. AI can plan and act faster than people can understand, interrupt or reverse.
Example use cases in financial services	Automating repeatable back-office tasks	Scoring, classifying and segmenting customers or transactions	Analysing messy data such as images, text, voice or behaviour	Drafting, summarising, coding and generating synthetic media	Executing multi-step workflows across systems

Summary: new layers of AI capabilities – and risks

The 2019 answer still broadly holds: ordinary automation follows rules written by humans; modern AI learns patterns from examples.

Longer-standing forms of 'rule-based automation' follow instructions explicitly specified in advance – if this happens, do that. Modern AI, built on machine learning, is different. The system learns patterns from data and uses those patterns to make a form of prediction.

The early days of machine learning

But what we call 'AI' covers a wide range of technologies that look very different in practice. At the time of the 2019 predecessor to this report, the main focus was what might be called 'classical machine learning'. This used established statistical techniques and algorithms (such as decision trees, support vector machines, regression models, clustering, and so on) to detect patterns in data. In financial services, these techniques could be used wherever better pattern recognition or prediction could improve decisions – credit scoring, fraud detection, customer segmentation, pricing, risk modelling, and much more.

The underlying ideas were not new. Many had been understood for decades. But they had become practical because, by the 2010s, the necessary ingredients had finally arrived at scale: large datasets and cheaper computing power.

A more powerful form of machine learning, known as 'deep learning', was also becoming more prominent. Deep learning uses neural networks – mathematical systems loosely inspired by the structure of the brain – to detect complex patterns in large and messy datasets.

This makes it especially promising for tasks involving freeform text, images, voice, and other forms of unstructured data. In financial services, this can mean matching faces and documents in identity checks. Or reviewing images and text in insurance claims. Or detecting subtle patterns across long sequences of transactions. However, the extra power comes with a cost. The model can produce impressive results, but is often impossible in practice to peer inside the 'black box' and understand precisely how it got there. Deep learning also tends to require much more data and compute than classical methods.

The dawn of Gen AI

Public awareness of AI changed dramatically when LLMs and Gen AI became widely accessible in late 2022. Earlier forms of machine (and deep) learning usually sat behind the scenes. They produced structured outputs – a fraud alert, a risk classification, a pricing recommendation. Gen AI was different. It produced unstructured, human-readable outputs – most visibly at first in the form of fluent text; and increasingly in images, audio, video and computer code.

With LLMs, two shifts mattered most:

- First, they could ingest messy inputs and return coherent answers in fluent language.
- Second, unlike earlier AI systems that were usually designed for a specific use case, LLMs felt general-purpose. The same tool and interface could be used to draft a letter, explain a contract clause, turn meeting notes into an action plan, write code, and hold a coherent discussion about both trends in banking fraud and the history of Test cricket. Applications could still be customised, but the base technology was no longer confined to one narrowly defined task.

At their core, LLMs are still machine-learning systems. They are trained on enormous quantities of text and learn statistical patterns in language. When prompted, they generate an answer piece by piece – roughly speaking, by predicting the next token, then the next, and so on. This is an oversimplification, but it captures the essential point: the system is not simply retrieving a verified answer from a store of facts. It is generating a plausible continuation of the prompt based on patterns it has learned (though later tuning, retrieval tools and safeguards can make its answers more accurate and useful).

Machine persuasiveness – right or wrong

Earlier AI systems were often wrong, of course. A structured output could look precise and still be mistaken, and users could still place too much confidence in it. But generative AI adds something different. It adds persuasion. It can explain, justify and contextualise its output in fluent language. The system can even reassure users by explaining itself – while giving an explanation that is incomplete, misleading, or simply a plausible-sounding story about how the answer was produced.[3]

This feeds into the now-familiar problem of hallucinations – where LLMs produce fabricated statements, often in the same confident tone as the rest of the answer. These errors may not appear as obvious nonsense; they can be embedded inside an otherwise accurate and helpful response. And this is not simply a bug waiting to be fixed.[4] Hallucinations arise from the same probabilistic machinery that makes LLMs useful – their ability to generate plausible language in open-ended situations. Safeguards can reduce the risk, but not guarantee reliability.

In financial services, the response to this might be that LLMs will not make final decisions; they will only support humans with insights. However:

Humans are prone to 'automation bias[5]*'*: We defer to machine outputs, especially when they sound fluent, confident and convenient.

- *Even without blind faith, people are busy, tired and under pressure.* "This looks good enough" is a very human response to a persuasive answer at the end of a long day.
- *The output may not stop with that user.* A mistaken interpretation of policy wording might be copied into a claims response; a fabricated regulatory citation might find its way into a compliance note. The danger is not only that a person believes the error. It is that the organisation absorbs it.

And hallucination is only part of the Gen AI shift[6]:

- *Gen AI lowers the barrier to using AI:* Natural-language and low/no-code tools allow non-specialists to build workflows and automate tasks. That brings opportunities, but also means AI can spread through organisations faster than governance can follow.
- *Gen AI extends beyond text.* It can be used to create images, audio, video and synthetic identities, which has direct implications for financial services. 'Deepfakes' can undermine voice authentication, video KYC, customer verification, fraud controls, and more. Which means that some of the basic evidence financial firms rely on – documents, voices, faces, instructions – becomes less reliable.
- *Gen AI makes AI feel ordinary.* Earlier AI often sat behind specialist systems. Gen AI appears as a chatbot, writing assistant, meeting summariser or spreadsheet helper. That makes it easier for employees and customers to bring AI into financial services informally, outside approved systems.
- *Gen AI is a supply chain – not just a model.* Behind a fluent chatbot answer sits a wider supply chain, which the Financial Stability Board helpfully lays out in five layers[7]: hardware, computing infrastructure, training data, pre-trained foundation models, and under-facing applications. This matters, because each layer brings different dependencies and vulnerabilities. A financial institution may interact with a customer-facing tool, but the real risk may sit deeper in the stack – in the cloud provider, the model developer, the data pipeline, or the hardware on which the whole system depends.

The future: Agentic AI – and beyond

Much of the AI hype has now moved to agentic AI: systems that do not just respond, but *act*. Given a goal, they can work through the steps, including using tools, querying data, updating records, and interacting with other systems.

Agentic AI is not a singular technological breakthrough in the way GenAI was. Before ChatGPT, most people had no intuitive frame for what it would feel like to ask a machine to write, summarise, code or converse fluently. Agentic AI feels more like the natural next step: once AI can generate useful outputs, the obvious question is whether it can also act on them. Whether or not it has GenAI's immediate "wow" factor, its implications could be as consequential. Financial services is full of processes made up of small, linked tasks: triaging fraud or AML alerts, gathering customer information, checking policy rules, drafting responses, updating records, and routing unusual cases for review. Agentic AI shifts the frontier from assisting with individual tasks to coordinating and executing workflows across systems.

The hard part is the plumbing. Gen AI could spread through ad hoc use: paste an email into a chatbot, summarise a document, etc. Agents need access to enterprise systems, data, permissions, audit trails, and workflows. This is where the risk changes. Gen AI raised the problem of whether AI outputs could be trusted. Agentic AI asks what happens when those outputs are connected to action, with a degree of autonomy.[8] A bad output might update a record, trigger a workflow, send a customer message, escalate a case, or initiate a transaction.

With older rule-based automation, an error could be repeated at great speed, but it was the same error following the same path, explicitly coded by humans. Agentic AI can carry a mistake through a chain of decisions, tools and actions – adapting as it goes. In other words, the risk is not just from automated execution of a bad rule; it is automated *judgement*. In a March 2026 report, the UK's Competition and Markets Authority warned: "Autonomy for agents increases the consequences of errors, may heighten risks of manipulation and loss of consumer agency, and could lead to worse overall outcomes for consumers… AI agents raise new questions about transparency, incentives and accountability and whether the current tools and frameworks that protect consumers are fit for purpose".[9]

Can we predict what will come next?

In 2019, almost no one has an intuitive feel for how quickly LLMs would enter ordinary work. The next shift may come from better agentic systems, cheaper compute, new model architectures, quantum computing, robotics, or something less visible today. This report does not try to assess the prospect of artificial general intelligence (AGI)[10] – though it is not unreasonable to ask where the current trajectory might eventually lead. The more immediate point is that AI risk is not static. Each leap changes not only what the technology is capable of – but also what can go wrong.

Risk drivers: The CSFI's 2019 predecessor to this report identified three characteristics of AI/machine learning that help to explain how it is different from longer-standing forms of automation, and why it may bring about new kinds of risks. They are still relevant today, but need expanding for a world of Gen AI and increasingly agentic systems.

Opacity and complexity	**Erosion of human control**	**Changing incentive structures**	**Embeddedness and dependency**	**Generativity and false fluency**
Generally speaking, the more powerful the model, the more difficult it is to understand, explain, validate and challenge.	Humans remain formally accountable, but may be increasingly distant from how decisions are generated, shaped or executed.	AI rewards speed, scale, data capture and automation, creating pressure to move faster than governance can follow.	AI is increasingly built into tools, vendors, infrastructure and workflows, creating hidden reliance and shared failure points.	AI can be widely used to produce persuasive text, images, audio, video and code that may be misleading or fabricated
Increasing in the pre-Gen AI era, and have further intensified since			**More recent priorities**	

How is AI used in financial services?

While AI use cases in financial services are not the main focus of this report, it is worth (briefly) setting out some context here. Not least, because risk depends on adoption: how widely AI is being used, and how deeply financial institutions are embedding it into their activities.

Two recent surveys are especially useful on this front. The Bank of England and FCA's 2024 survey[11] on AI in UK financial services received 118 responses from financial firms, reporting aggregated results. The 2026 Global AI in Financial Services Report[12], produced by the Cambridge Centre for Alternative Finance and partners, surveyed 628 organisations across 151 countries[13]. Together, they show:

- *AI adoption is already mainstream.* The BoE/FCA survey found that 75% of respondent firms were already using AI – up from 58% in the 2022 survey – with a further 10% planning to use it over the next three years (i.e., by 2027).
- *AI is expected to move deeper into firms.* Respondents to the BOE/FCA survey expected the median number of use cases to rise from 9 to 21 over three years, with nearly a quarter expecting more than 50.
- *The pace of adoption of Gen AI is striking.* The Cambridge survey found that, among industry respondents, adoption of generative AI had reached 71% – close to classical machine learning at 75% – despite GenAI only gaining widespread traction since 2022. The report links this uptake to lower barriers to adoption than traditional machine learning methods. It also found that agentic AI is already in active adoption among 52% of industry respondents.
- *Current deployment of AI is still mostly operational.* The Cambridge survey found that the top four industry AI use cases were related to operations: software development, data visualisation, data and knowledge management, and process automation. Similarly, the Bank/FCA survey found that operations and IT accounted for the largest share of reported AI use cases.

The takeaway, then, is that AI use in financial services – including Gen AI – is widespread, and much of this use is about improving execution, rather than (necessarily) reinventing financial services. This implies that the main risk – at least for now – is less about firms handing core decisions to autonomous systems. It is that AI is entering more workflows, faster, and often in the background functions that keep financial institutions running.

The use-capability gap

The fact that AI can *theoretically* be used to do something doesn't mean it will be in practice. In the Cambridge survey, 55% of industry respondents and 63% of regulators said it was difficult to measure the value of AI deployment. Firms still need to establish a worthwhile return on investment. Even then, the barrier to adoption is broader than the business case alone: firms also need the data, infrastructure, governance and skills to use AI safely and effectively.

One impediment to adoption that regulated firms often raise is fear of regulatory sanctions and fines if they deploy AI in ways that later prove to be unsafe or unfair. As discussed later in this report, the UK has espoused a 'principles-based' approach to regulating AI in financial services – one which avoids enshrining prescriptive rules around a fast-changing technology, and treats AI as part of existing obligations. The argument is that this should support flexibility, and avoid suppressing innovation too early.

There is a persuasive case for this type of approach. Yet it is also criticised for, in effect, transferring interpretive risk to firms. If the practical message feels like: "you decide how the rules apply, and we will tell you afterwards whether you got it wrong," cautious firms may slow down, or avoid higher-value use cases. The risk is that a regulatory approach designed not to supress innovation can have quite the opposite effect – if firms feel like regulators are passing the buck.

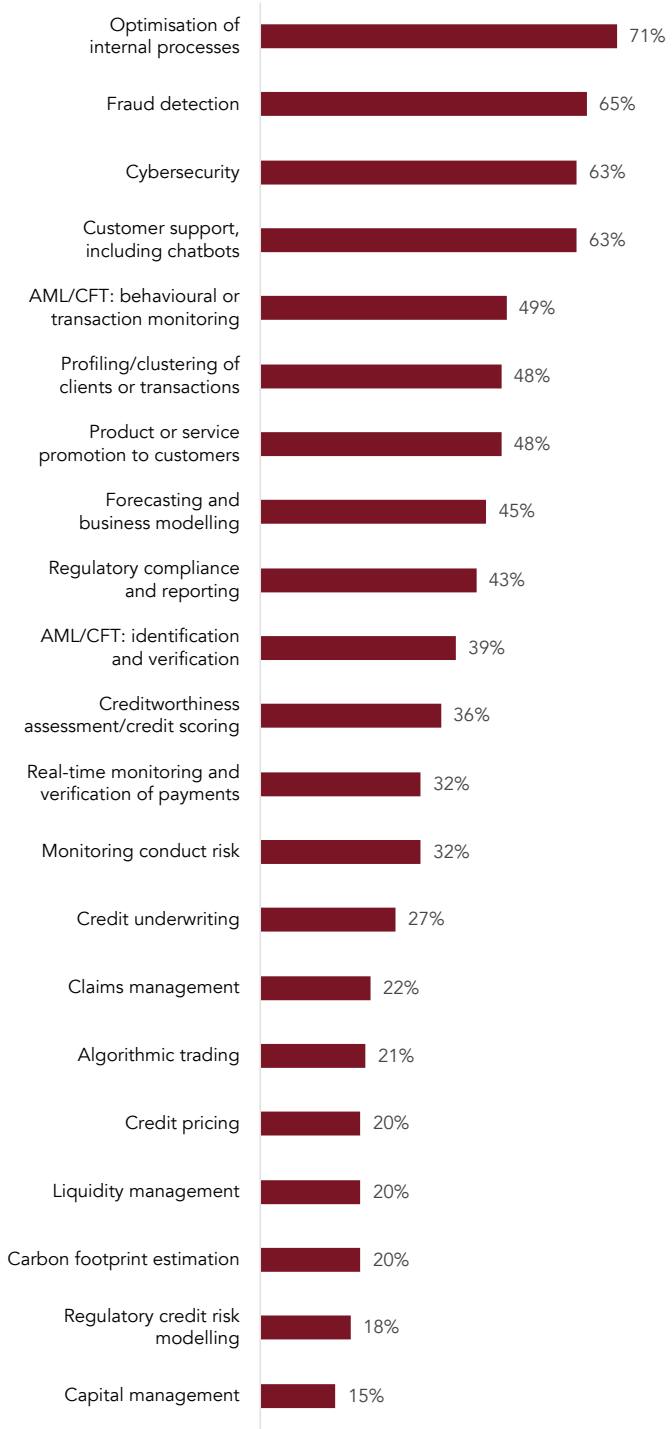
But regulatory caution is just one side of the story. AI can create strong incentives to move quickly in markets where advantages compound and the winners take most. Whatever the balance is now, it is reasonable to ask how long it will hold. Does the fear of sanction outweigh the fear of falling behind – or, indeed, of being unable to compete at all in a market reshaped by AI?

Nor does adoption necessarily stay neatly contained. Firms might begin with low-risk, internal use cases, gain confidence, and then over time drift toward customer-facing, revenue-generating or control-function applications.[14]

The AI use-capability gap, then, cuts both ways – regulatory uncertainty can make firms overly cautious, but market pressure can also push them too far, too fast.

Use case examples

The chart below shows the percentage of respondents in the Bank of England/FCA survey who, as of 2024, were either using AI or planning to use it (within the next three years) for different purposes. The table on the right provides an illustrative example of what each of the ten most common use cases could involve.



Top 10 use cases Examples of AI applications in financial services
Optimisation of internal processes A tool that extracts information from customer documents and pre-populates internal review forms.
Fraud detection A model that detects unusual payment patterns and flags likely fraud before a transaction is approved.
Cybercrime A tool that monitors network activity and flags suspicious access attempts for investigation.
Customer support, including chatbots A chatbot that answers routine account questions and escalates complex or sensitive cases to staff.
AML/CFT: behavioural or transaction monitoring A tool that identifies transaction patterns consistent with possible money laundering.
Profiling/clustering of clients or transactions A model that groups customers or transactions with similar behaviours for risk analysis.
Product or service promotion to customers A recommendation tool that suggests relevant financial products based on customer behaviour or profile.
Forecasting and business modelling A model that forecasts demand, losses or cashflow to support business planning.
Regulatory compliance and reporting An AI assistant that reviews documents against internal rules before regulatory reporting.
AML/CFT: identification and verification An AI tool that compares identity documents and biometric checks during digital onboarding.

* Illustrative examples were selected by this report's author and were not provided by the Bank of England/FCA. AML/CFT is 'anti-money laundering and combating the financing of terrorism'.

A moving baseline

The BoE/FCA 2024 survey remains a strong baseline, but it should be read as a snapshot of reported use cases from a fast-moving market. It captured a financial sector where AI was already widely adopted, but still largely used for analytics, automation, and human-supervised decision support. Since then, the frontier has moved further toward LLM-enabled productivity and workflow tools – often still pilots or assistive systems, but already changing how work gets done.

As of mid-2026, AI is not routinely making high-stakes regulated decisions on its own in the financial services sector. The near-term story is less about unconstrained autonomy than AI doing more of the work around human-accountable decisions – with more agentic tools beginning to emerge.

The opportunities

There are compelling reasons for financial institutions to adopt AI. Used well, it can lower costs, improve service, strengthen risk management, widen access, and help organisations make better use of information. This report focuses on risks – but the implication is that AI's opportunities are significant enough to make its risks worth taking seriously.

The 2019 predecessor to this report identified six broad categories of opportunity. They are still very relevant today, but the table below shows how they have evolved in the Gen AI era.

Alongside these firm-level opportunities, for many years much of the excitement around AI has been that it can help democratise financial services: widening access to credit, savings, insurance, and advice, and providing lower-cost alternatives to services that had traditionally been too expensive for many customers. The language of 'democratisation' is easy to overuse, but the underlying promise is real.

What is newer is the idea that AI could become a more continuous form of support. Rather than simply automating a task, answering a question, or personalising a product, AI could help customers and staff navigate financial decisions over time. It might, for example, alert a customer that their regular expenses are beginning to outpace income, or help them prepare questions before speaking to an adviser.

Overview: Opportunities AI brings to financial services, 2019 → 2026

2019 framing of opportunity	How the opportunity has expanded in the Gen AI era
Faster, cheaper operations, such as streamlining processes, automating regulatory compliance, and analysing incoming data in real time.	**From automating tasks to reshaping workflows.** AI can now support whole processes, not just isolated tasks – for example, reading a customer file, drafting a summary and routing it for review.
Curtailing or eliminating human error, while freeing human talent to purse more creative or higher-value work, such as by reducing manual mistakes in repetitive tasks.	**From replacing repetitive work to assisting judgement-heavy work.** AI can help staff with more complex work, such as drafting an investment note or checking a compliance response before a human makes the final call.
Always-on digital services, such as chatbots, digital channels and automated customer interactions available outside normal business hours.	**From 24/7 access to more natural digital conversations.** Customer support can move beyond scripted chatbots, with AI helping customers explain a problem, understand options and complete routine steps in plain language.
Customised products and advice, tailored to the unique characteristics of customers or clients, such as personalised pricing, investment decisions, and insurance underwriting.	**From targeting customers to guiding them personally.** AI can personalise not only what a customer is offered, but how it is explained – for example, showing how different mortgage choices affect monthly payments.
More accurate predictions and better decisions, such as better credit scoring, underwriting, fraud detection, and risk management using large datasets.	**From better predictions to better-informed judgement.** AI can help translate model outputs into action, such as explaining why a fraud alert matters and what an investigator should check next.
Insights from new types of data and new markets, such as using alternative, unstructured or non-traditional data to reach underserved customers.	**From analysing more data to making messy information usable.** Gen AI can turn unstructured material into usable knowledge, such as summarising claims notes, customer calls, or policy documents at scale.

In the future, AI agents may begin to complete multi-step tasks on users' behalf, such as continually comparing insurance policies, alerting the customer when a better option is available and, with permission, even switching provider automatically. If that type of functionality becomes widespread, digital finance starts to look less like a set of products and channels, and more like AI-enabled support as an ongoing service.

The risks of not using AI

This report's focus on the risks posed by AI does not imply that avoiding AI is risk-free. Quite the opposite is true, in fact. Non-use is not neutrality.

The most obvious risk of not using AI is the missed opportunity, such as the categories of benefit laid out in the table. But there is more to it than that. In some areas, AI use is increasingly necessary not to capture the upside, but to protect against the downside – in an operating environment that is changing rapidly.

Cybersecurity is the clearest example. Bad actors are using AI to generate more convincing scams, automate attacks, create synthetic identities, and probe sensitive systems from countless different directions, thousands of times per day. 'Not using AI' as a defensive measure, here, is often simply not an option. A similar logic applies to competition. If rivals use AI to lower their costs or gain market share, firms that do not adapt could struggle to compete, even to survive.

There is also a question of visibility. As financial activity becomes faster and more data-rich, some risks may emerge too quickly or too subtly for human monitoring alone. That is why many financial institutions are embedding AI into risk management functions. Used well, it can help firms spot patterns that would otherwise be missed, and manage risks across large, complex systems.

There is an equivalent point for supervisors. If firms are using AI, and if AI-enabled risks can move quickly through markets and infrastructure, regulators cannot rely solely on traditional supervisory tools. They need the ability to monitor firms, detect weak signals, analyse large volumes of data, and identify where the next financial shock may be building.[15] In short, AI can create risks, but it can also be used to anticipate them.

The baseline has changed

Financial institutions and regulators are not deciding whether AI enters the financial system from a standing start. In important respects, it is already there. Gen AI is not like earlier, specialist machine learning tools that sat mainly inside institutions. It is a general-purpose technology that customers and employees can access directly in their everyday lives.

That changes the terms of the choice. Customers are already using general purpose AI tools to compare products, draft complaint letters, interpret investment options, and seek financial guidance – whether or not regulated firms provide those services themselves. Employees may be using AI to summarise documents, prepare analysis, write code, or draft communications – whether or not those tools have been formally approved.

In other words, choosing not to deploy AI does not necessarily mean avoiding AI-related risk. If regulated firms move too slowly, customers will likely turn to less reliable tools outside the regulatory perimeter[16], and staff may use public systems in ways that create data, compliance or conduct risks. Similarly, if supervisors do not develop AI capability, they may be trying to oversee a financial system increasingly shaped by AI with tools built for a slower, less automated environment.

A quote from the predecessor to this report is as relevant now as it was in 2019: *"I think it is very important that as we look at these new systems, we should not be comparing them to perfection, but rather to how we do things today."* The cat is now very much out of the bag. AI is already part of the ecosystem in which financial services operate – used by customers, employees, criminals, vendors and competitors alike.

> *"The risk is that regulators and Government make perfect the enemy of good – and whilst we take too long to deliver this, more people are accessing inappropriate or unregulated advice."*
>
> Head of regulatory affairs, fintech sector

The risks: outcomes

How AI affects customers and society

This report maps AI risk across three levels: Outcomes, Operating Environment, and System. *Outcomes* are the human end of the risk chain – where AI's effects are most directly felt. They concern who gets access, who is excluded, what customers can understand or challenge, and whether AI-generated outputs are reliable enough to support real financial decisions.

Bias and social harm

The Risk	Caused by…	Leads to…	What has changed?
AI turns narrow optimisation into unfair or socially damaging outcomes.	Biased data, biased model design, and commercial objectives that reward prediction, targeting or profit over fairness.	Some groups benefiting while others are excluded, overcharged or steered toward worse outcomes – weakening fairness, inclusion and the social value of finance.	Gen AI and more advanced behavioural modelling make it easier to personalise prices, offers, and customer interactions – creating more ways to identify vulnerability, steer choices, and exclude less profitable customers.

The uncomfortable tension: *Fairness to whom, and for what reasons?*

Erosion of transparency and trust

The Risk	Caused by…	Leads to…	What has changed?
AI makes finance feel more intrusive – with decisions made by a black box that watches more, explains less, and is harder to challenge.	Opaque models, complex inputs, unclear accountability, and AI use cases that feel creepy, unfair or poorly explained.	Customers, firms and regulators struggling to understand why a decision was made, how to challenge it, and who is responsible when it goes wrong.	Gen AI can make answers sound clear without making decisions explainable; while as AI becomes embedded in everyday software and vendor tools, it can shape outcomes without being recognised or challenged.

The uncomfortable tension: *'Defensible' is not the same as 'meaningful'.*

Misleading outputs and hallucinations

The Risk	Caused by…	Leads to…	What has changed?
AI can be wrong in convincing ways – generating plausible falsehoods or turning incomplete data into overconfident conclusions.	Models working from incomplete data, weak assumptions, gen AI hallucinations, and systems that generate likely answers rather than verified truths.	False or misleading outputs being relied on in ways that affect customers, firms, or financial decision-making.	Gen AI introduces a new failure mode: it can simply make things up, while presenting them in polished, confident language that makes errors harder to spot. Persuasiveness makes automation bias harder to resist.

The uncomfortable tension: *Does 'mostly right' break the use case?*

Bias and social harm

THE CORE TRADE-OFF:

Inclusion through sharper prediction, versus exclusion through sharper discrimination.

AI excels at optimisation, and financial services is full of things that can be optimised: pricing, targeting, approval, risk assessment, etc. In principle, better prediction is appealing on a societal level because it can widen financial inclusion, for example, by identifying creditworthy people who lack traditional data or fit awkwardly into older models. But the system optimises for what is measured and rewarded. If a model is designed to maximise profitability, repayment, engagement, or risk accuracy, it may follow its instructions brilliantly while producing outcomes that are less fair or inclusive. This means that in practice, AI can create winners and losers among customers. Some enjoy better pricing, access, or service, while others face worse terms or exclusion – and those who lose out are often among those least able to absorb the loss.

In the Pre-Gen AI Era: This risk was largely framed as a problem of bias in data and models. Machine learning systems trained on historical data could reproduce and amplify historical inequalities. If past lending patterns were biased, the model would learn those patterns and if the data were incomplete or skewed, the outputs would be, too. There was also a wider concern that more granular risk assessment might undermine traditional ideas of fairness in finance – for example, in insurance, by making it easier to differentiate between customers rather than pool risk. At the core of this was the question: If finance becomes too good at identifying the profitable customer, what happens to everyone else?

More recently: Generative AI, richer behavioural data, and more granular targeting have raised the prospect not just of consumers being sorted, but *steered*. The longer standing (and still very relevant) fear is discrimination. A newer one is extraction: where AI identifies vulnerability – conditions like financial stress, confusion, desperation, even cognitive fatigue – and optimises around it. In other words, it is about AI being used to guide customers towards outcomes that suit the provider better than themselves. Meanwhile, the use of LLMs to shape messaging makes this steering more personalised and persuasive, and finely tunes its timing. The troubling possibility is a financial system that is *systematically* able to extract value from weakness.

If personalisation becomes extraction, the consequences can go beyond direct consumer exploitation:

- **Accuracy can undermine dismantle solidarity.** In insurance and credit, more granular risk assessment implies the possibility of less pooling, less cross-subsidy, and less room for imperfect but socially valuable judgement.[17] The system can become worse at absorbing difference in favour of pricing it.
- **A new thin-file underclass can emerge.**[18] People with sparse data, irregular lives, weak digital footprints, or a reluctance to share information may be treated as poor risks not because they are genuinely unsafe to serve, but because they are hard for the system to read. The real dividing line may not be creditworthy versus not creditworthy – but optimisable versus non-optimisable.
- **A two-tier system of protection can develop.**[19] Higher-value or better-served customers may get AI wrapped in human oversight and easier recourse. Meanwhile, mass-market or more vulnerable customers are left with raw automation and a greater burden to spot problems themselves.

A problem is that 'consent' is often a weak safeguard, here. In practice, what often happens is that terms are bundled, bargaining power is weak, and most people cannot meaningfully choose how their data are used. As one department head at a large financial institution put it to me, this lack of meaningful choice has "a profound societal impact, because it means privacy becomes something only the rich can afford. Part of allowing that box to read your data is giving that data to a company that can then do what it likes with it".

Why technical fixes are often not enough

'Blind' algorithms are largely an illusion. Removing race, gender, or other protected characteristics from a dataset does not make an AI system neutral if it can reconstruct them through proxies. That is what AI excels at: it finds correlations in apparently innocuous variables and turns them into socially meaningful distinctions. The system does not need to be intentionally prejudiced to reproduce structural inequalities with great efficiency.

A concern expressed often by people designing these systems is a kind of legal *Catch-22*.[20] To test whether an AI system harms particular groups, firms often need access to exactly the protected data – race, gender, disability – that privacy and equality frameworks make difficult to collect or use. This means that the technical route to proving fairness can collide directly with the legal framework meant to protect it. Firms may be told not to discriminate, while being constrained in how they can even measure whether discrimination is occurring.

The absence of intent does not make the harm less real

Harm often arises without malicious intent. It can emerge from many individually rational design choices interacting in ways that systematically disadvantage certain people. There may be no memo saying "exploit vulnerability" – only a system doing exactly what it was built to do.

And better individual decisions can still worsen aggregate outcomes. More accurate targeting may help firms optimise each interaction, but across the system it can mean more precise nudging into debt, weaker risk pooling, and greater hidden fragility. While each decision may look rational on its own terms, together, they can create harm that no single decision seems responsible for.

The uncomfortable tension: *Fairness to whom, and for what reasons?*

Fairness is not a single thing, and trying to pin down a definition is fraught with difficulty. Is it about equal treatment? Equal error rates? Equal access? Pricing that reflects risk? Risk pooling? Cross-subsidy? Consumer autonomy? Protection of the vulnerable? These cannot all be satisfied at once.[21] So every AI system deployed in a high-stakes financial context is, whether explicitly or not, making a subjective choice about which version of fairness to privilege.

And that leads to the more uncomfortable point: some of what we call 'bias' is not just a technical defect to be fixed. Sometimes it is the visible edge of a deeper political question: what do we want financial services to be for? If we want inclusion and risk pooling, then we may have to accept some of the frictions and inefficiencies that AI is explicitly designed to eliminate. That is why this risk is so thorny. *The tension is not just between fairness and bias. It is between finance as an optimisation problem and finance as a social institution.*

What makes this more awkward is that the law may not be built to see the problem clearly. AI-driven harm does not always map neatly onto protected categories or familiar concepts of discrimination. Models can disadvantage people in ways that are real in their effects but hard to name in legal terms. So even when harm is widely felt, it may not fit the boxes the law is most comfortable enforcing. The implication is that AI is not just testing our models of fairness; it may also be exposing the limits of the legal categories we rely on to defend it.

> *"The risk is the proliferation of a technology which engenders trust, but can abuse that trust. People may trust these things, particularly for financial advice, and then get taken advantage of. What's to stop someone asking, "I need some advice, what should I do with my money?" and being guided towards products and services which may not be appropriate for them? Commercial incentives mean these things could happen – and that is a major systemic societal risk."*
>
> Non-executive director & chartered actuary

Erosion of transparency and trust

THE CORE TRADE-OF:

Performance through complexity, versus trust lost if decisions cannot be explained.

Much of what makes AI attractive in finance – complexity, scale, and the ability to find patterns humans would miss – are the very things that make it harder to understand, explain, and challenge. That is the bargain, essentially baked into the technology. As systems become more powerful, they tend to loosen the link between decision and explanation in any form a human can meaningfully grasp. The trade-off, then, is more than just performance versus transparency (i.e., can we identify which inputs affect which outputs?). It is between greater machine capability and human legibility. The importance of financial services to people's lives makes it one of the worst places to discover that those are not the same thing.

In the Pre-Gen AI Era: This risk was largely about 'black box' decision-making. Machine learning models could outperform simpler rule-based systems, but at the cost of making decisions that were harder to explain to customers – and to regulators. This was particularly true with the most opaque 'deep learning' techniques. It mattered most in customer-facing contexts such as credit and insurance provision, fraud detection, or AML, where firms are typically expected to justify why a person has been denied a product, flagged for suspicion, or priced a certain way. More than that, opacity weakens accountability: if nobody can explain why the model behaved as it did, who is responsible when it gets something wrong? Trust weakens when decisions start to feel inscrutable, arbitrary, or detached from ordinary human reasoning.

More recently: Opacity is more than just a problem of hidden logic – it is now also a problem of simulated understanding. Gen AI has made systems fluent and conversational, which creates the impression of transparency. In some ways, that is more dangerous than the old black box. A silent model at least advertises its opacity. A chatbot that sounds articulate can trick users into thinking they understand both the answer and the system behind it. At the same time, AI is being embedded into more parts of firms' operations, often invisibly, through vendors, tools, interfaces and, at the frontier, agentic systems that can act across workflows. That makes accountability harder to locate. A thornier problem than opacity within a single model is opacity across a growing web of interacting systems.

Explainability ≠ explanation

The key distinction, here, is that a model can be interpretable in a technical sense without offering a real explanation. A bank may be able to show which variables influenced an outcome, or how a model's weights shifted, without giving the customer anything that feels intelligible or actionable. What the customer usually wants to know is *"was this fair?"*, and *"what would I need to change to get a different result?"* The problem is that AI can often answer the first question only weakly, and the second one not at all.

Furthermore, different stakeholders need different types of explanations.[22] Regulators want to understand whether the system is broadly controlled and non-discriminatory. Customers want a plain-language reason for their outcome. Senior managers want enough understanding to sign off without being blindsided later. Explanations can undermine trust when firms give one audience the kind of explanation meant for another – which can feel evasive or patronising.

AI logic can feel arbitrary – and creepy

Even a clear explanation may not be reassuring if the underlying logic feels alien. Trust also breaks down when decisions are shaped by inputs that feel bizarre or intrusive. If financial outcomes depend on proxies like someone's smartphone device behaviour, browsing patterns, or other seemingly irrelevant signals, customers are left with a deeply uncomfortable question: what hidden rules am I being judged by? Decisions can start to feel less like ordinary financial judgment and more like a hidden behavioural scoring regime. Even if the correlations are statistically predictive, the logic can feel illegitimate and creepy: as though we are being constantly tracked, and arbitrary details of our behaviour could shape life-changing decisions.

Accountability can become performative

Financial services need accountability to sustain trust, but AI makes it harder to locate. Even if there is a named person who is formally responsible for a decision, the reality may be that they are accountable on paper for a system they neither fully understand nor meaningfully control. The question, then, is: *if things go wrong, do we actually have accountability, or just a scapegoat?* The problem becomes harder when the system spans model providers, vendors, infrastructure firms and regulated entities. The risk is that each party can (plausibly) say they are only responsible for one small part of the chain – meaning nobody truly owns the whole.

Trust erosion is more than a conduct issue

The consequences can go beyond complaints or reputational damage: a collapse in trust can have has liquidity consequences. For example, a widely publicised AI failure or a wave of complaints about arbitrary treatment can erode confidence extremely quickly. And once moving money is frictionless and instant, mistrust can travel faster than any formal response. What begins as a complaint-handling issue can rapidly become a stability issue.

… And even the evidence is getting shakier

There is also an important practical point here. AI is weakening some of the foundations on which trust rests. If voice can be cloned, video can be faked, and recorded interactions are no longer straightforward evidence, then many of the basic mechanisms through which firms establish trust and resolve disputes become less trustworthy. In other words, proof itself becomes less reliable.

The uncomfortable tension: *'Defensible' is not the same as 'meaningful'.*

The financial services ecosystem needs explanations that people can trust, challenge, and use to assign accountability. Yet it is not clear that AI can provide explanations that are both technically faithful and meaningful to humans. A model can be interpretable in principle, yet there may be no single explanation that is simultaneously true to what actually happened, intelligible to the person affected, and useful for holding someone to account.

That creates a temptation to settle for something weaker: something just *defensible*. It's good enough to reassure a supervisor, close a complaint, or satisfy a governance process. A chart is produced. A human signs off. A senior manager is named. The explanation exists, so the system appears governable. But if that explanation does not genuinely help someone understand or change the outcome, then the real problem has not been resolved.

Yet what is the alternative? There is a danger in assuming that more interpretability automatically makes things better. If firms are pushed away from more capable but opaque systems toward simpler models that are easier to explain, they may end up with the worst of both worlds: weaker outcomes and only superficial comfort.

A 'glass box' that works badly but looks governable may be worse than a black box that works well but is honestly difficult to understand. The – perhaps inevitable – tension is that the systems that perform best are never explainable in a way humans find satisfying, while the systems that feel most explainable are too simple to handle the most complicated decisions.

> *"It is likely given existing regulation that any implementation of AI will have to supply an explanatory audit trail for any decisions. The danger is that this gives the appearance of transparency without actually surfacing the correlations that drove the AI's decision."*
>
> Partner, Life & Pensions Consulting

Misleading outputs and hallucinations

THE CORE TRADE-OFF:

Useful answers in messy situations, versus confident answers that may be false.

AI is particularly attractive when the task at hand requires something like 'judgement' – to interpret messy information, fill gaps, and produce usable answers in situations where reality is only partially observed. But unlike the risk of over-optimisation, the danger here is not that AI is 'too good' at its job. It is that the system may be wrong – because its data are incomplete, its model is poorly matched to the problem, or, in the case of Gen AI, it can produce fluent but fabricated outputs.

In the Pre-Gen AI Era: The main concern was that AI systems could produce outputs that looked rigorous but were unreliable for deeper structural reasons. In many financial contexts, models only ever see part of the picture. For example, a bank's fraud or AML system sees only the data available to that institution, not the full world it is trying to judge. The problem was (and still is) not just that models were sometimes wrong, but that they could be wrong in ways that were hard to spot in advance. The model may be missing important information without anyone knowing exactly what is missing; inputs may look valid but be misleading; and patterns that held in the past may fail to generalise. The deeper risk is *epistemic overconfidence*: trusting technically sophisticated systems to make judgements that rest on a weak grip on reality.

More recently: Older machine learning systems often failed in contained ways, such as a wrong risk score. Gen AI 'hallucinations' feel fundamentally different. Outputs sound coherent and authoritative even when they are false. This matters in financial services (as well as more widely) because the danger is not just being wrong, but being wrong in a way that is persuasive enough to be trusted. The problem becomes sharper when firms use AI for tasks where 'mostly correct' is not good enough. If a human still has to check everything because the downside of error is too high, much of the efficiency benefit starts to disappear. As well as introducing a new kind of 'persuasive wrongness', Gen AI has exposed a deeper question about which use cases can bear that risk at all.

A different kind of danger

There is (of course) nothing new about errors, computer made or human. What makes the risk from LLMs different is that the error does not arrive looking uncertain or fragile – it looks confident. The problem is that AI wrongness can masquerade as knowledge.

In fact, the word 'hallucination' may be slightly misleading because it makes the model seem like it is briefly glitching. In reality, the system is doing what it was built to do: producing statistically plausible sequences without real understanding. Some AI observers prefer to use the (perhaps more honest) term "confabulation",[23] which shifts the framing from 'bug in the system' to 'part of the architecture'.

A broader concern than outright fabrication is that outputs can vary from one run to the next, even when the task appears unchanged. One senior data scientist told me: "Deterministic and repeatable is a requirement in our industry. You put in the same input, and you get the same output. The issue with Gen AI is that it's non-deterministic and non-repeatable. So you put in the same inputs twice, and you could get a different answer. It's not just about trusting the algorithm for accuracy, it's about trusting the algorithm for repeatability."

The real harm is often what happens next

The most serious failures may not come when someone follows one bad output directly. They come when that output is treated as a valid building block and fed into the next step. The risk is workflow contamination: wrongness spreading because it is embedded in apparently routine processes.

Hallucination is only one form of wrongness

LLM hallucinations are the newer and more visible failure mode. But the older problem is as relevant as ever: AI can be wrong because it is trying to infer too much from too little. A model may be founded on incomplete data, unstable proxies, or assumptions that only look reasonable until you ask how they would really be tested. In that sense, hallucination is only the most conspicuous expression of something wider: systems being asked to make claims their evidence base cannot fully support.

The problem can be a weak premise

Sometimes the issue is not that a system has slipped up, but that it was never in a strong position to know the thing it is being asked to judge. In fraud, AML,

or behavioural prediction, the model may only ever see a fragment of the world it is trying to interpret.[24] A system can look sophisticated, well-governed, and statistically clean, while still leaning on a weak premise. This is epistemic overreach hidden beneath a layer of statistical polish.

The fragility often sits in the plumbing

Another reason this risk is hard to manage is that the weakness may not sit neatly in the model itself. It may be in the data being fed in, the way the system is set up, the handoff between one tool and another, or the point where a human is expected to intervene. The model may appear well grounded in theory and still mislead in practice, with the fragility hidden inside the wider system.

Governance may be supervising the wrong kind of failure

Financial regulation and internal governance were designed for familiar failures: bad disclosures, weak controls, flawed models, and so on. They were never built for a world in which automated systems can produce confident, fluent falsehoods at scale (because until very recently, that world simply did not exist). In other words, hallucinations do not fit neatly into the categories firms and regulators are used to supervising. The problem, then, is deeper than 'the system can be wrong'. It is that the rulebook was written to govern a different species of wrongness.

The uncomfortable tension: *Does 'mostly right' break the use case?*

LLMs are useful because they do not just retrieve fixed answers. They generate responses to messy and open-ended human prompts – which is why they feel flexible, helpful, and intelligent. But that flexibility comes from a probabilistic architecture that is trying to produce a plausible next response – not to stay silent unless it 'knows' something to be true. Truthfulness is desirable, and firms can push models toward it, but it is not the thing the system most fundamentally optimises for. In that sense, hallucination is not a random defect bolted onto an otherwise truth-seeking machine. It is a by-product of using a generative system to respond fluently where the input is ambiguous, incomplete, or only partly answerable.

The problem is that many financial tasks do not merely require something plausible or helpful. They require something correct on the points that matter – and those points are not always easy to specify in advance. A summary can sound accurate while omitting the one point that changes the decision. A compliance draft can be polished while misstating the clause that matters most. And so on. Yet if a human must check everything, much of the efficiency case disappears.[25]

The tension arises when we try to use systems built to generate plausible language as if they were systems built to produce truth. For some use cases, that may be acceptable. But many parts of finance run on precision, trust, and tiny details that matter. In those contexts, it may well be the wrong bargain: firms could spend huge amounts of money and governance effort trying to make AI look safe and reliable, when the real problem is that the task requires a level of certainty the technology simply cannot provide.

> *"The huge challenge is that you get models to a point where they are, say, 85% trustable and believable, but you can't identify ahead of time the 15% where it will fail. That's where your risk sits. And there seem very much to be diminishing returns in trying to get that 85% higher."*
>
> Head of Analytics & Technology, Reinsurance

How AI changes the conditions firms operate in

The second level of risks is the *Operating Environment:* the conditions in which financial institutions adopt, use, and control AI. Here, the risks come less from one bad model decision and more from the world AI creates around firms – stronger attackers, faster adoption pressures, and a growing gap between what AI can do and what people can understand or control.

Cyber threats and privacy breaches

The Risk	Caused by…	Leads to…	What has changed?
AI vastly expands both the amount of sensitive data in use and the ability of attackers to exploit it.	Firms using more sensitive data in AI systems, attackers using AI to target them, and staff or vendors exposing information through tools outside firms' control.	Severe breaches, fraud, data misuse and disruption – with potentially existential damage to firms' reputations and customer trust.	Gen AI makes scams, deepfakes, phishing and impersonation cheaper, more convincing and easier for more actors to scale — while everyday AI tools create new ways for sensitive data to leak.

The uncomfortable tension: *An arms race with no finish line.*

Human talent and knowledge gaps

The Risk	Caused by…	Leads to…	What has changed?
AI increases machine capability faster than human understanding, leaving firms less able to build, use, govern or challenge it properly.	Shortages of people who can bridge AI, financial risk and regulation; weak AI literacy across the wider workforce; and leaders held accountable for systems they may not fully understand.	Hollow governance, over-reliance on AI outputs, and humans rubber-stamping systems they do not have the knowledge, authority or confidence to challenge.	Gen AI and low/no-code tools make AI easier to use without making it easier to understand – widening the gap between access and real human control.

The uncomfortable tension: *Is 'human in the loop' a comforting illusion?*

Rushed or poorly managed adoption

The Risk	Caused by…	Leads to…	What has changed?
AI's perceived transformative potential pushes firms to deploy before they truly understand the use case, controls, or consequences.	Competitive pressure, hype, fear of missing out, weak deployment discipline, and AI tools that feel easy to use but are hard to govern.	Poorly tested systems, inappropriate use cases, weak safeguards, 'shadow AI', and growing reliance on tools that were never properly assessed.	Gen AI feels like an everyday tool rather than a formal deployment, making it easier for low-risk experiments to drift into high-stakes reliance before governance catches up.

The uncomfortable tension: *Can firms show AI is safe before it is used in the real world?*

Cyber threats and privacy breaches

THE CORE TRADE-OFF:

More data-driven defence, versus more data-rich targets.

AI tends to be most powerful when it moves from the periphery of organisations to their nervous systems: when it is fed more data, plugged into more systems, and woven into live decisions. But that is also what makes it dangerous. Firms need more data to make AI useful, yet that same data becomes more valuable to steal, easier to misuse, and harder to control, especially once shared across tools and vendors. And, as attackers embrace AI, defending against them often requires adopting even more AI, which raises the stakes again. The difference is that attackers only have to succeed once; defenders have to succeed every time.

In the Pre-Gen AI Era: In the 2010s, cyber threats were seen as an inevitable part of digitisation and reliance on the internet. Financial institutions were becoming more dependent on technology vendors. There were growing risks related to interference with systems, data theft, insider misuse, and vendor failure. As the decade progressed, machine learning sharpened those risks because it rewarded data collection and centralisation: the more information firms could pull together, the more useful their models became. That meant more sensitive material in circulation, and more vulnerable targets for bad actors. It also changed the shape of the risk. Privacy was no longer just about theft or leaks, but about what could be inferred once enough data were stitched together.

More recently: The longer-standing risks are as relevant as ever, but Gen AI has added an entirely new layer. It lowers the skill threshold for attackers, making scams, phishing and impersonation cheaper, more personalised, and more convincing. It also expands the range of what can be faked. Anyone can now use simple prompts to generate more convincing emails, documents, application materials, etc. More sophisticated forms of fraud are also becoming more accessible. Voice cloning, deepfake video, and synthetic identities threaten the evidence firms use to verify who someone is and whether an instruction is genuine. That makes the risk feel fundamentally different from traditional cybercrime. More than systems being breached or data being stolen, trust itself has become easier to counterfeit.

The attack surface now includes language

We used to imagine cyber risk as code being used to attack code – technically proficient hackers probing networks and firewalls. That picture is now too narrow. With Gen AI, ordinary language becomes part of the attack surface. For example, prompt injection (the use of hidden instructions that trick an AI system into ignoring its rules or doing something unintended) allows attackers to exploit the same conversational interface that customers or employees use for help, not just the software underneath.[26]

Fraud is becoming experimental

Bad actors can test many different messages, identities, application variants, or attack routes, and learn quickly from what gets through. This enables a wholly different tempo from traditional fraud – with cheap experiments, rapid feedback, and constant iteration. In some ways, the attacker starts to look less like a burglar and more like a product team optimising a funnel.

A survey respondent from the risk and governance sectors said: "Firms are catastrophically under-aware of how easily AI-enhanced fraud can penetrate their *human* gatekeeping. This is a culture risk: disaffected staff, in particular, present a huge attack surface for the new generation of behaviourally-aware 'social engineer' hackers."

… and identity is becoming easier to fake

Deepfake audio and video, and synthetic identities (fake identities assembled from a mixture of real and fabricated personal information), are not just 'more fraud'. They weaken the basic evidence financial institutions use to decide who someone is. If the evidence firms are accustomed to relying on becomes shakier as proof, KYC (for example) starts to fundamentally lose some of its credibility.[27]

Privacy risk is no longer just about leaks

The privacy story is subtler than database infiltration. AI makes it easier to infer sensitive information from fragments that appear innocuous enough on their own. This is the 'mosaic effect'[28]: data points that look harmless can reveal intimate financial or personal information when combined at scale. This means that

privacy risk is not just about 'was the file stolen?' It is also 'what can now be reconstructed, inferred or exposed once enough pieces are joined together?'

… And leaks can be unwitting

Cyber and privacy risk does not have to require malicious intent. Employees can expose sensitive material simply by putting it into public or poorly controlled AI tools.[29] Data can leave an origination simply because a stressed employee uses a convenient tool to summarise a client note or clean up a spreadsheet.

Old plumbing makes new AI risks messier

In practice, AI is rarely deployed in a neat, self-contained environment. It is layered onto ageing systems, vendor tools, data feeds, permissions structures, and manual workarounds. This means that attackers often do not need to compromise 'the AI model' itself. They can exploit the seams: a weak integration, a poorly controlled data feed, an unmanaged update, or an access permission nobody has reviewed for years.

As an AI expert and actuary put it to me: "Big companies have more data and access to more powerful systems. So, in theory they can benefit from AI more than smaller firms or startups who don't have as much data. However, big companies are also known for being slow. Their data may have been organised according to models designed at the end of the 90s or in the 2000s – and they keep collecting the data following these models".

The attack can be against trust itself

An AI-enabled attack can do serious damage by targeting trust directly, even without breaching a bank's systems. A fake video of a CEO appearing to admit liquidity problems, or AI-generated messages claiming withdrawals are being blocked, can create enough uncertainty for customers to act. When banking is digital and withdrawals are almost frictionless, people do not need to fully believe the rumour; they only need to conclude that moving their money is safer than waiting to find out for sure.

The uncomfortable tension: *An arms race with no finish line*

Using AI in cyber-defence is no longer really a choice. As attackers use AI to generate more convincing scams, probe systems faster, impersonate customers, and automate experimentation, financial institutions cannot realistically respond with legacy controls. They are pushed toward sophisticated defensive tools: more automated monitoring, behavioural analytics, data sharing, real-time detection and response, and so on. That is unavoidable. But the logic is that the answer to AI-enabled risk is often more AI.

For financial providers, this creates a grim asymmetry. Attackers can fail cheaply and repeatedly while defenders have to be right continuously. One successful breach can undo countless successful blocks. And the defensive response itself expands the risk surface. There are more data, models, vendors, and integrations that can fail or be manipulated. The system becomes more complex and dependent. It is harder to control.

The disquieting tension, here, is that there does not seem to be any stable endpoint. This is not a risk that gets 'solved'. It is an escalation dynamic. Each improvement in defence changes attacker behaviour, and each improvement in attack forces more defensive action. How much money, complexity, customer friction, and management attention can firms prudently devote to chasing a moving target that never disappears?

> *"As an industry, we like to talk about agility and agile building – about how fast we are. But think about the real process that goes into a product, or a campaign at a bank. I have an idea. What happens next? I run it up the flagpole, check with my manager, flesh it out. Then it goes to legal and compliance, then to the data science team to build it, then back through legal and guardrails, then to a customer for testing. From idea to implementation, that takes at least a year, probably longer.*
>
> *Now you're a fraudster. You have an idea. What happens next? Ten minutes later, if even that, you're at your computer testing, implementing and refining. And if one bank figures it out, the attacker can move to another bank. If banking as a sector figures it out, they can move to tax offices or insurers. Their investment keeps paying dividends even after they get caught."*
>
> Global Head of Data Science, Cybersecurity

Human talent and knowledge gaps

THE CORE TRADE-OF:

Powerful tools for more people, versus fewer people able to challenge them.

The degree to which AI ultimately augments or replaces human work aside, it's clear that human implementation is one of the most critical factors in whether AI systems succeed or fail. And because AI can fail in serious and unpredictable ways, firms need people who can understand, challenge, and override it. But as applications become more sophisticated and embedded, meaningful human oversight becomes harder to sustain. The more powerful the system, the greater the need for human control – yet the harder it may be for humans to exercise that control meaningfully.

In the Pre-Gen AI Era: As financial institutions adopted increasingly complex machine learning models, the pool of people who could both understand the technical side and grasp the wider business implications – regulatory, conduct, operational, and so on – was thin. Finding talent that could build AI well was hard enough; finding domain experts who could challenge them in context, and explain risks to senior management, was perhaps even harder. Data scientists often lacked business expertise, while compliance, risk and business teams lacked the understanding to interrogate data scientists' work. The result was a governance gap. Firms could have controls on paper while still lacking the people able to ask the basic questions: *is this the right problem for AI, is the model using the right data, and do we understand the risks well enough to deploy it?*

More recently: Five years ago, most people had little direct experience of using AI interfaces, even if AI systems were already affecting them behind the scenes. Today, almost everyone who reads this has encountered an AI chatbot, and generative AI and low or no-code tools have widened access dramatically. More employees can now use AI through ordinary language, without writing code or really knowing how the systems work. There are obvious opportunities from this. But it also widens the gap between who can use (and depend on) AI, and who can genuinely challenge or govern it. Gen AI is especially risky because its fluency can create a false sense of comprehension. Agentic AI could sharpen this further, as the difficulty moves from supervising outputs to supervising actions. The easier the interface becomes, the easier it is to overlook how little users may understand about the underlying system.

Firms may be adopting AI faster than they understand it

The Bank of England and FCA's 2024 survey showed how AI adoption in financial services is spreading quickly – across more firms, more functions, and more types of application, including generative AI. (See the earlier section in this report: *How is AI used in financial services?*) But the survey also contained a striking admission: almost half of respondent firms (46%) said they had only a "partial understanding" of the AI technologies they use – compared to a third (34%) reporting "complete understanding". This gap was especially acute when it came to third-party models, where firms rely on systems they did not build and cannot fully inspect. Put another way, AI is becoming embedded in financial services before many firms fully understand what they are embedding.

The rarest skill is translation

The gap in understanding is about more than 'not enough technical AI experts'. The harder problem is finding people who can translate between worlds: model design into financial regulation, statistical performance into customer outcomes, technical controls into board-level accountability, and so forth. A data scientist may understand the model, but not the consumer duty implications. A compliance officer may understand the rulebook, but not the ways in which the model can fail. The crucial blind spot often sits between disciplines.

Fluency creates false comfort

Gen AI makes shortcomings in understanding more treacherous because it feels easy to use. An LLM that explains itself smoothly gives users the impression that both the tool and the output are understood. It may even produce an explanation of its reasoning, but that explanation can itself be misleading (or simply a plausible story after the fact).

Governance can be hollow

It is difficult to solve these problems through governance processes alone. A firm can have governance committees, risk policies, sign-off processes, and escalation routes, and still lack the

human capacity to challenge what the system is doing. The same problem applies at the top of organisations. Senior managers and boards may be expected to approve and take responsibility for systems they cannot meaningfully scrutinise. Even if "I didn't understand it" does not hold up as a defence, it may still be true. The important question is not about whether a process exists, but whether anyone in that process understands enough to ask the right question, spot the weak assumption, or stop the deployment.

… And 'hidden AI' and handovers create blind spots

Firms may not even be fully aware where AI is being used inside their organisation. It increasingly arrives embedded inside ordinary tools: vendor software, analytics platforms, customer-service systems, and productivity applications. Even when firms know a tool uses AI, knowledge can still get lost at the handover points – from developer to vendor, vendor to firm, and firm to user. Assumptions and caveats disappear; limitations are forgotten.

AI can hollow out the expertise it still needs

If AI takes over tasks such as routine analysis, underwriting, compliance drafting, or risk assessment, it may remove the training ground where humans develop judgement. Junior staff learn by doing the repetitive, foundational work. If that work disappears, how will the next generation of leaders acquire the instincts needed to challenge machine-made decisions?

A survey respondent, a compliance director at a large bank, said: "If we become over-reliant on AI, who will be maintaining the knowledge base and expertise to correctly challenge and overrule it? We have already seen examples that are so convincing only subject matter experts have been able to identify the hallucinations. In the drive for efficiency arising from AI, there is a real risk that this expertise gets sidelined".

The uncomfortable tension: *Is 'human in the loop' a comforting illusion?*

In a regulated system like finance, there is a strong temptation to say that we can absorb the risks of AI so long as there is a 'human in the loop' – where a person reviews the AI's output before acting. But what does that really mean? If this person lacks the expertise to challenge the system, if they cannot keep up with the volume of work (a 'tired person ticking a box'), or if they are psychologically inclined to defer to fluent AI outputs (a cognitive bias known as *automation bias*), then they may be present without being meaningfully in control.

Yet if we insist on genuinely expert human review at every critical point, a lot of the promised efficiency gains from AI begin to disappear. Often, the stronger the business case for AI, the weaker the business case for intensive human scrutiny becomes. This is a tension that is difficult to escape. The technology is supposed to reduce dependence on scarce expert judgement, but may require that judgement to be used safely.

That is why this risk is not just a matter of training more people or hiring more specialists. AI changes the relationship between human judgement and machine output. If the system takes over the work that humans have used historically to build expertise, fewer humans may be left with the judgement needed to challenge it. The real test is not whether a human is 'in the loop', but whether that human has the knowledge, time, authority, and independence to change the outcome. Otherwise, it feels less like control than a story we tell ourselves to make AI decisions feel governable.

> *"We have to learn how to use AI, recognising both strengths and weaknesses. Perhaps the biggest challenge will be the loss of knowledge as analyst and junior jobs are replaced by AI taking away the traditional means by which people learn and progress to more senior roles. This is especially true for risk quantification, model building and validation. If we rely on AI to build risk models - who will judge the quality or validity of the results if they have not learnt by building models themselves?"*
>
> Chief Executive, Investments and Lending Risk Manager

Rushed or poorly managed adoption

THE CORE TRADE-OFF:

Competitive advantage through speed, versus avoidable harm through haste.

The AI opportunity is not merely about marginal improvements to efficiency. It promises something qualitatively different from previous technologies, potentially transforming every area of business: faster decisions, lower costs, better customer service, new products. Firms also face pressure to use AI as proof to investors, boards, and customers that they are innovative and not falling behind. That creates pressure to act quickly, especially when competitors appear to be moving first. The risk is that firms move to deploy AI before they have worked out where there is a genuine need for the technology, what can go wrong, and what controls are needed.

In the Pre-Gen AI Era: This report's 2019 predecessor warned that firms risked implementing AI "for its own sake" – when simpler and cheaper alternatives could have worked just as well. At the time, one expert in financial regulation told us: "Lots of businesses will say 'we need to have an ML [machine learning] expert'… so they'll hire a ML expert and instruct them to 'go and find problems". Part of the problem was hype: firms expecting AI to deliver transformative results while underestimating the data, infrastructure, and specialist support needed to make those results achievable. But there was also pressure from the other direction. Firms afraid of falling behind could rush to deploy AI before use cases, safeguards, testing, and long-term maintenance had been properly worked through. And unclear accountability could leave nobody fully responsible for how the system behaved once it had been deployed.

More recently: AI no longer has to enter a firm through a formal model-development process. It can arrive through chatbots, AI assistants, vendor software, everyday productivity tools, or informal staff use of public AI systems. The pressure to adopt is also no longer just top-down. Unlike in the pre-Gen AI era, staff and managers can often see, in plain language, how these tools make their own work faster or easier. It is easy to become frustrated when official channels block uses that feel obvious and low risk. On the surface, building a prompt or AI-assisted workflow can feel closer to, for example, creating a custom spreadsheet than deploying a regulated model. The risk is uncontrolled diffusion: AI spreading through workflows before firms have mapped where it is being used, what data it touches, who is accountable, or – more fundamentally – whether it is genuinely appropriate for the task.

This risk is not simply recklessness

The pressure to adopt AI is largely rational. But the technology's very real promise is often interwoven with hype, fear of missing out, and a need to show stakeholders you are not being left behind. The danger is that AI adoption becomes largely about signalling momentum.

In many digital businesses, a failed experiment means a clumsy product release. In financial services the stakes are often much higher. A failed experiment can affect credit decisions, savings, insurance, advice, complaints, disclosures, or regulatory obligations. As one risk professional put it to me: "It feels like we're in a technology race. But are we being thoughtful enough? Is this the right application? Too often, the logic is simply that we want to optimise, become more efficient, become more productive, so we throw more technology at the problem".

In December 2024, a panel of experts' report to the G7 warned that: "Institutions acting in a highly competitive environment may prioritize the speed of deployment and short-term performance gains over thorough testing and risk assessment of their AI systems"[30]. In other words, the danger is that the competitive environment rewards corner-cutting – and AI risk becomes a collective-action problem.

The report added: "When there are externalities, competition provides strong incentives to race to the bottom and forces companies to ignore their external effects on safety and system stability to stay ahead of their competitors, unless they are forced to do so by their regulators".

Gen AI feels like an everyday tool, not a deployment

A few years ago, AI in financial institutions felt like something specialist and remote from the rest of the business – the sort of thing that required technical teams and formal approval. Gen AI often feels much more mundane: a prompt, a document summary, a spreadsheet-like workaround, a small workflow hack. That makes it easier for employees to see the benefit and harder for firms to make the risk feel real.

'Shadow AI' muddies the waters

The problem for firms that respond simply by blocking (or heavily restricting) official use of AI tools is that their employees may then be tempted to use public tools, personal accounts or informal workarounds. The firm may then lose track of what data are being entered, what outputs are being relied on, and where AI is shaping work in the background.

Low-risk uses can drift into high-stakes reliance

Gen AI adoption typically starts with relatively safe, mundane tasks, away from sensitive areas: summarisation, drafting, coding, support, etc.

But commercial pressure rarely stops there. And nor should it, necessarily. If AI is genuinely useful, firms will naturally look for ways to move it closer to customers and higher value tasks.

The danger is how it gets there – if it is incremental creep rather than a deliberate decision. For example, if a pilot becomes useful, staff start depending on it, managers see productivity gains, and suddenly the system is part of the workflow. The important question is not just whether the original experiment was approved, but whether the firm noticed when that transformed into reliance – and reassessed risks, controls, and accountability accordingly.

…And procurement can import risk before deployment begins

Particularly when third-party vendors are involved, firms can make poor AI adoption decisions before a tool is ever switched on. The risk starts in procurement – buying systems without enough clarity about how they work, how data are handled, how they change over time, and how much control the firm will really have if something goes wrong.

The uncomfortable tension: *Does 'mostly right' break the use case?*

Firms often cannot know where AI is genuinely useful until they let people use it. Its value is discovered in messy workflows, with real users, real data, and real institutional constraints – it is not just academic. But finance is not a forgiving place to learn by trial and error. A bad experiment can harm customers or fall foul of regulatory obligations. The implication is that some degree of real-world exposure may be necessary to learn how to use AI well, but that exposure is itself a source of risk.

Regulatory sandboxes can help by creating a supervised space for firms to test AI with real users and regulatory engagement. But they do not eliminate the underlying tension. A sandbox can reduce the risks of experimentation, but it cannot fully reproduce what happens once AI is scaled beyond a controlled test and absorbed into ordinary business activity.

And restraint does not eliminate the problem. If firms move too fast, they risk weak controls, bad use cases, and systems being implemented before they are properly governed. But if they move too slowly, staff may find their own routes through public tools and workarounds. The choice, then, is often not simply 'AI versus no AI.' It is governed adoption versus unofficial adoption.

That is what makes this risk especially thorny. Firms may know they are not fully ready and still feel they cannot simply stand still. The pressure is more than hype (though there is plenty of that). It is competitive anxiety and the fear of appearing technologically obsolete.

> *"A risk is that you just trust AI too much. You implement an AI approach, maybe driven by marketing pressure, because other companies are already saying they have AI. You feel you should go for it too. But you are not able to assess whether the implementation is actually good. Eventually, it can become unprofitable, or fail to grasp all the patterns in the data it was trained on, and then you end up in trouble because you adjusted your risk appetite based on AI-driven decisions."*
>
> Senior actuary in financial services, and data science expert

How AI risks scale and propagate

The third level in this risk framework is the *System:* how AI risks scale, interact, and propagate across the wider financial system. These risks arise when regulation struggles to keep up with adoption, markets reward speed and scale over resilience, and firms converge on the same models, infrastructure, or providers – making failures easier to spread and harder to contain.

Regulatory gaps and fragmentation

The Risk	Caused by…	Leads to...	What has changed?
AI evolves faster than rulemakers' ability to govern it effectively, while older regulatory frameworks struggle to fit what AI is becoming.	AI models and deployment practices evolving quickly, fragmented rulebooks, limited supervisory capacity, and systems that span firms, vendors and jurisdictions.	Gaps in oversight, inconsistent rules and expectations, and firms being accountable for systems they may not fully control, intensifying the risk that harmful uses of AI go unchecked or are unevenly controlled.	Generative AI and large third-party models push risk outside traditional regulatory categories – leaving supervisors to govern systems whose design, infrastructure and ownership may sit beyond the regulated firm.

The uncomfortable tension: *Is regulation reactive by necessity?*

Market distortion and skewed incentives

The Risk	Caused by…	Leads to...	What has changed?
AI can turn competition into a race for scale, speed and control of its critical capabilities – even when firms are cautious of breaking the rules.	AI's economics and regulatory uncertainty favour firms with the scale, compliance capacity, and bargaining power to move, while others hesitate.	A market where firms race to keep up, smaller players depend on dominant providers, and safety can lose out to speed.	Gen AI can make markets look crowded at the surface while power concentrates underneath – in models, cloud, compute, data and distribution.

The uncomfortable tension: *What if competition creates concentration?*

Shared dependencies and contagion

The Risk	Caused by…	Leads to...	What has changed?
AI can tie firms to the same hidden machinery – creating common failure points that spread shocks across the system.	Firms independently choosing the same models, data sources, cloud providers, software tools and third-party AI services because they are efficient, powerful or hard to replace.	Outages, bad data, model errors or flawed updates hitting multiple firms at once – and spreading faster than human controls can contain.	Gen AI can embed the same models, data sources and providers across many workflows, creating uncomfortable echoes of the financial crisis: risk looks dispersed at the surface while common dependencies build underneath.

The uncomfortable tension: *Parallels with the financial crisis*

Regulatory gaps and fragmentation

THE CORE TRADE-OFF:

Innovation through flexible rules, versus gaps where old frameworks do not fit.

Regulators are trying to govern a technology that is fundamentally reshaping financial services, while it also creates risks across every dimension of the industry. Move too slowly, and AI systems may become embedded in the ecosystem before supervisors understand their effects. Move too aggressively, and regulation may suppress innovation, favour incumbents, or lock in assumptions about a technology that is still changing. The temptation is to rely on existing rules and let the framework evolve naturally. But AI does not sit neatly inside existing regulatory boxes. It cuts across jurisdictions, risk categories and industry boundaries. Crucial parts of the AI ecosystem may also sit outside the direct reach of financial regulators. Part of the thorniness of the regulatory question is that financial institutions depend on model developers, cloud providers, data vendors and software platforms they have little control over.

In the Pre-Gen AI Era: The regulatory challenge was taking shape around two linked concerns. The first was safety. Early machine-learning systems raised concerns about data quality, bias, opacity, accountability and consumer harm. If models learned from incomplete or skewed data, they could reproduce unfair outcomes at scale; as deep learning advanced, the 'black box' problem made those outcomes harder to supervise. The second was innovation. Regulators were asking how to protect consumers without suppressing useful experimentation, especially in a market where AI could favour scale, blur the line between financial and non-financial firms, and create scope for regulatory arbitrage across sectors and jurisdictions. By the early 2020s, these questions were moving into formal regulatory processes, such as the UK's AI Public-Private Forum[31] and the EU's 2021 proposal for an AI Act.

More recently: The EU AI Act had to be adapted during negotiation to address emerging Gen AI capabilities. That is almost the perfect regulatory parable: the world's most important AI law was still being written when the technology changed under its feet.[32] The deeper issue is that Gen AI makes it harder to identify what, exactly, is being regulated. It is a shifting mix of models, prompts, data sources, vendor tools and human workflows. At the same time, key capabilities increasingly sit with large technology providers outside the traditional financial regulatory perimeter – raising questions about who should be on the hook for what. The challenge, now, is supervising an ecosystem with boundaries and owners that are increasingly blurred, particularly as agentic AI pushes systems towards greater autonomy and less direct human control.

The problem is more than regulatory 'lag'

The concern that regulation is failing to 'keep up' with AI capabilities is very valid, but it is one part of a larger story. AI does not map neatly onto the categories financial regulation is used to supervising. For example, is a Gen AI customer assistant a conduct issue, a model risk issue, an outsourcing issue, a data protection issue, or all of the above? When a technology cuts across risk categories, responsibility can fragment even when nobody is obviously asleep at the wheel.

Firms know that existing obligations apply, but not exactly what those obligations require for a new AI system. That is one of the dangers of 'technology-neutral' regulation. Applying rules focusing on outcomes that apply regardless of the technology used seems (and perhaps is) sensible in principle. But it leaves firms to do much of the interpretive work themselves. The risk is that technology-neutrality turns into passing the buck and a transfer of uncertainty – from supervisor to firm, and perhaps ultimately, from firm to customer.

Accountability without control?

Financial regulators can hold institutions responsible for how AI is used, but many of the crucial levers may sit elsewhere – with model developers, cloud providers, data vendors or software platforms. The 2026 Treasury Committee report on AI captures this posture neatly. It says firms are "on the hook" for AI-related harm, and that "I did not understand it" is not a defence. Yet if a firm cannot fully inspect the model, control the training data, veto an update, audit a vendor tool or roll back a change, what kind of control does it really have?

Accountability may be clear on paper while practical control is much thinner.

The broader problem is that calling this simply 'third-party risk' can understate the issue. Responsibility can become so widely distributed that, when something goes wrong, everyone can point to someone else. Who is legally responsible after the event matters, but the sharper question in practice is: *who has the power to act before harm spreads?*

Regulators may be structurally outgunned

AI supervision is, perhaps inevitably, an asymmetric contest. The firms and technology providers building these systems have most of the data, money, and human talent. Regulators are often trying to understand them from the outside, with thinner teams and budgets. In other words, supervisors are expected to govern systems they can only partially see, often learning their significance after the fact, armed with a fraction of the resources of the sector they are expected to hold to account.

Recent evidence from the FCA and Bank of England to the House of Commons Treasury Committee illustrates what this means in practice. Regulators in the UK are not proposing to inspect AI models directly; they are relying on outcomes-based supervision, firm accountability, and existing frameworks.

For example, the FCA's executive director for payments and digital finance said[33]: "I would not expect my supervisors to be looking at the individual model. They are not coders and they will not go into that, but under interventions such as the consumer duty, firms are expected to design products with a target market in mind, with outcomes in mind, and they should know what to expect."

That may well be the only realistic approach. But it makes clear that supervision will rely on the premise that firms' own governance will notice when the machine goes wrong – without regulators having to peer inside.

Authorities may not speak with one voice

As well as across borders (see next page), regulatory fragmentation can exist within a single jurisdiction. Because AI is so multifaceted, it attracts attention from almost every part of the public sector. For example, we are seeing finance ministries prioritise growth through innovation, conduct regulators consumer protection, central banks systemic stability, competition authorities market power, and so on. None is wrong, but together they can pull in different directions. If public authorities do not share a settled view of what kind of risk AI is, it is not surprising when the rules feel fragmented.

The uncomfortable tension: *Is regulation reactive by necessity?*

AI regulation may always be reactive – and perhaps it has to be. Regulators cannot fully supervise a technology before they know how it will be used, how firms will depend on it, where it will fail, and what new forms will replace the current ones. But by the time those facts are visible, the market may already have reorganised around them. Products are built, vendors selected, workflows embedded, and dependencies normalised.

There is also a conflict between clarity and adaptability. Firms want clear rules. But highly specific rules can become brittle when change is fast. Flexible principles age better, and they sound like common sense: do not panic, do not over-legislate, let existing frameworks adapt. But flexibility can turn into fog. If the framework is too unclear to give firms confidence, yet too thin to manage serious risk, it creates the worst of both worlds. Cautious firms underinvest or delay, while aggressive firms press ahead.

AI exposes a regulatory paradox. The more precisely regulators define the rules, the faster those rules may age; the more they rely on broad principles, the more uncertainty they shift onto firms and markets. There may be no neat resolution.

And the pace of change is startling. The Gen AI boom is the warning shot: public LLMs went from non-existence to board-level priority in a remarkably short period. Whether the next leap is agentic AI or something not yet visible, regulation may be structurally condemned to govern the last generation of technology while the next one is already arriving.

Regulatory fragmentation in practice: different models, different risks

Major jurisdictions are trying to solve the same problem in perceptively different ways. For example, the UK is leaning on existing rules and supervisory experimentation; the EU has built a formal AI rulebook; and the US is taking an explicit deregulatory approach while leaving oversight largely to existing laws.

UK: supervisory experimentation without a single AI rulebook

Who is setting the agenda? The FCA, Bank of England and PRA are setting the financial services agenda, with Parliament applying pressure from the side. Sandboxes and live-testing schemes – controlled environments where firms can test AI applications with regulatory engagement – are key tools, alongside the Bank/FCA AI Consortium, which is designed to gather industry input on AI use in finance. The FCA expanded its AI Live Testing programme in April 2026 to include eight more financial institutions.[34]

Regulatory posture: The UK broadly aims to be principles-based. Existing regimes such as Consumer Duty, SM&CR, operational resilience, outsourcing and model risk are expected to do much of the work. The posture is: let firms innovate, engage closely with supervisors, and learn through testing.

EU: The landmark AI Act, wrestling with implementation

Who is setting the agenda? The European Commission, Parliament and Council have set the agenda through the AI Act. Unlike the UK, the EU has chosen to legislate horizontally across sectors, including financial services. This makes the EU's approach more formal and rules-based than the UK's – and also, potentially, more influential if global firms treat it as the benchmark.

Regulatory posture: The AI Act entered into force in 2024 and treats certain financial uses – such as creditworthiness assessment – as high-risk, bringing stricter obligations around governance, transparency and human oversight. The harder part is implementation. Through the Digital Omnibus process, EU institutions have moved toward delaying high-risk obligations, with implementation dates expected in 2027 and 2028.[35]

US: A deregulatory tone, with financial risk handled through old levers

Who is setting the agenda? At the federal level, the White House is now setting a sharp deregulatory tone. in January 2025, President Trump revoked President Biden's earlier AI executive order – which had focused on safety testing, agency standards and risk mitigation – and later pushed a national AI framework aimed at removing barriers to US AI leadership and limiting state-level rules.

Regulatory posture: There is no single US AI regime for financial services. Instead, regulators are applying their existing powers – from consumer protection and fair lending to banking supervision, model risk, and fraud. In April 2026, the Federal Reserve, FDIC and OCC revised model-risk guidance to exclude generative and agentic AI, treating them as too novel and fast-moving for the existing framework and leaving them to broader governance controls.[36]

Market distortion and skewed incentives

THE CORE TRADE-OF:

Compounding advantage through scale, versus markets rewarding concentration.

AI both depends on scale and rewards it. Systems feed on enormous volumes of data and compute (as well as needing infrastructure and human talent) which naturally favours large firms and centralised providers. At the same time, AI also tends to create compounding advantages: better models attract more users, generate more data, improve faster, and make it harder for others to catch up. The upside is that scale can turn AI into a powerful, widely available layer of financial infrastructure. The danger is that the same forces concentrate power, raise barriers to entry, weaken diversity of approaches, and turn a small number of dominant providers into critical chokepoints.

In the Pre-Gen AI Era: AI seemed to be spurring competition in the short term, with fintechs and Big Tech firms challenging incumbents – but questions were being asked about whether that competition might later give way to winner-takes-all dynamics. The World Economic Forum predicted[37] that AI in financial services could lead to the "bifurcation of market structure", where scale-based institutions captured customers from mid-sized firms, but smaller, nimbler players might carve out niches for themselves. There was also growing commentary about the potential of machine learning creating 'data network effects' – a virtuous cycle of automated product improvements that may make it very difficult for others to effectively compete in the market. The issue, then, was not just 'big firms could get bigger'. It was that AI could reshape the competitive foundations of finance.

More recently: Concerns about the financial sector's dependence on third parties have always been on the radar, but they are now front and centre. Gen AI has intensified third-party concentration because the most capable systems depend on an infrastructure stack – compute, cloud, chips, data centres, energy – that only a small number of providers can supply at scale. And the dependency is not just deeper; it is also wider. Earlier AI tools were often narrow: a fraud model, a credit score, or an underwriting tool. Gen AI and agentic systems can become a general-purpose layer across multiple functions. The sting is that technology providers are often larger, more technically capable, and more powerful than the financial firms they serve. What happens when accountability stays with the firm – but control over transparency, data, updates and resilience sits with the tech provider?

When first-mover pressure meets regulatory caution

A key theme in the CSFI's 2019 report was that AI distorts financial institutions' incentive structures: that "the benefits to successful actors and the risks of getting left behind create powerful incentives for firms to collect data and implement AI solutions on a rapidly accelerated timeline." That is still true, but the dynamic is more complicated in regulated finance. Many firms are deeply nervous about breaching rules and attracting supervisory scrutiny.

But caution does not remove the competitive pressure; it changes its form. Regulatory uncertainty often favours large incumbents, who are more able to absorb the cost of legal advice, model governance, and vendor due diligence. Smaller firms may struggle to do this, leaving them more exposed if they move quickly, or more dependent on off-the-shelf tools and dominant technology providers. Concentration of markets, then, can come from a growing gap between firms with the capacity to manage the transition themselves (or at least on terms they can shape) and those forced to rent it.

Vendor concentration undermines substitutability

When it comes to technology providers, being large is not automatically the problem. The more practical question is whether financial firms can realistically switch away if the provider changes terms or the model, degrades its service, or suffers an outage. Concentration becomes dangerous when the service is critical and credible alternatives are scarce.

This matters both among technology providers and among the financial institutions that depend on them. The potential ramifications are familiar, but sharper in an AI context:

- *Barriers to entry:* smaller firms struggle to compete if the core AI infrastructure is controlled by a few players.
- *'Too-important-to-fail' dynamics:* critical AI providers could become so embedded in financial services that their failure or withdrawal would create wider instability.
- *Moral hazard:* if providers become essential infrastructure, markets may assume they will be supported or protected in a crisis.

Can regulators challenge providers the system depends on?

Dominant AI providers may also become difficult for policymakers and regulators to challenge if a handful of technology firms are seen as essential to growth, innovation and national competitiveness. For example, in the UK, the government recently said growth should sit "at the heart of regulators' remit", with a stronger duty to support business growth and proportionate, business-friendly regulation.[38]

The UK has also created a "Critical Third Parties regime", intended to give regulators powers over firms providing essential services to the financial sector, including AI and cloud providers (the rules came into effect from the start of 2025). But in its 2026 report on AI in financial services, the Treasury Committee was plainly frustrated that no firm had yet been designated, even after an Amazon Web Services (AWS) outage highlighted the financial sector's dependence on major cloud providers.

Recognising dependence, then, is not the same as constraining it. A risk is that regulation starts from the premise that these providers must be accommodated, rather than from a clear view of what the financial system needs from them.

The stranger possibility: collusion without a conspirator

Another more theoretical concern – but one which has been raised by the Bank of England[39] – is that AI could make markets less competitive without anyone explicitly setting out to collude. This could happen, for example, if AI systems designed to optimise for profit observe their rivals – and learn that reinforcing each other's behaviour is the best strategy. No one needs to instruct the system to collude; it may simply find that coordination-like behaviour is profitable. That sits awkwardly with a legal system built around human intent and communication, because machine optimisation may produce anti-competitive outcomes without either.

The uncomfortable tension: *What if competition creates concentration?*

AI may produce the appearance of competition while eroding some of the diversity that makes competition valuable.[40] On the surface, financial services can look full of choice, with lots of different service providers, vendors, and AI-enabled tools and apps. Yet underneath, many of them rely on the same infrastructure: foundation models, cloud providers, chips, data centres, and coding tools. In fact, the apparent paradox is that intense competition at the surface *produces* convergence underneath – many firms racing to adopt the same tools because not adopting them looks commercially irresponsible.

The uncomfortable reality is that this may not be a distortion of the market, but a consequence of the market working exactly as expected. AI typically works better at scale. The market may rationally converge on a few providers because they are the best (or at least because scale makes them look like the safest, cheapest option) – rather than because anyone is cheating. Can ordinary competition fix a market where the underlying economics reward concentration?

Banks, insurers and asset managers are regulated entities with formal duties to customers, supervisors and the stability of the financial system. But core AI capabilities are increasingly supplied by a handful of technology providers whose incentives and accountability are not organised around the safety of financial services. And critically, these providers are not simply keeping the lights on. They are supplying part of the decision-making engine that shapes essential services. Financial institutions may still own the customer relationship, while parts of the intelligence layer increasingly sit elsewhere.

Shared dependencies and contagion

THE CORE TRADE-OFF:

The compelling logic of common tools, versus the fragility of common failure points.

Financial institutions quite rationally gravitate toward the AI systems and suppliers that seem most capable, trusted, and easy to integrate. Few firms have the resources or appetite to build everything themselves. Using established providers is the common-sense choice. But when many firms make the same sensible decision, the system can become dangerously uniform. What looks responsible for one firm – using the best model, the best cloud, the best data feed, or the most trusted vendor – can make the whole system more fragile if everyone else does the same. The trade-off is individual efficiency versus system-wide resilience.

In the Pre-Gen AI Era: This risk was not created by machine learning. Older forms of automation also came with contagion risk: common exposures, shared infrastructure, outsourced critical services, herd behaviour, and so on. Most notoriously, the 2007–08 financial crisis was, in part, a lesson in how apparently diversified institutions can turn out to be exposed to the same assumptions and failure channels. What machine learning added was not just 'more automation', but a more complicated form of common dependency. The 2019 CSFI report pointed out that: "widespread AI use could lead to co-variance between previously uncorrelated systems". It warned that concentration among AI service providers could create systemically important institutions, 'too big to fail' dynamics, and 'too connected to fail' concerns, with isolated failures potentially cascading.

More recently: It is easy to think of AI contagion mainly through automated trading or flash crashes. While that is part of the story, the risk is broader. AI is becoming part of the plumbing of ordinary financial services – as relevant to a retail bank handling customer queries, fraud alerts, and complaints as it is to a trading desk. Gen AI sharpens the risk because it can become a shared interpretive layer across firms, from summarising information to shaping judgement. Even where the front-end tools look different, they may rest on similar model families and update cycles underneath. Agentic AI could potentially take this further: beyond generating outputs, systems could trigger actions, escalate issues, reroute workflows, or adjust exposures. At its core, the risk is that shared AI tools can turn common interpretation into common behaviour – and common behaviour into contagion.

Hidden common dependencies

In the previous section on 'market distortion and skewed incentives', concentration mattered because it could weaken competition and entrench dominant providers. Here, it matters because it creates common failure channels. The question goes from 'who has too much power?' to 'what happens if too many firms depend on the same thing?'

A financial institution can diversify its vendors without necessarily diversifying its dependencies. For example, three different fintech partners may look like three independent providers, but if they all rely on, for example, the same cloud provider or LLM family, the institution has only diversified the front end. That can create false comfort for firms (and supervisors). On paper, the firm has spread its risk across multiple suppliers. In reality, those suppliers may all depend on the same underlying machinery.

An even thornier problem than third-party risk is known as 'nth-party risk', where the problem is pushed further down the chain. A firm may know its direct vendor, but not the vendor's vendor, the model provider behind that, or the data pipelines feeding it. The systemic weak point can be several contractual layers away from the regulated firm. The danger of relying on ordinary due diligence, here, is that it can obscure the true dependency map.

The risk that shared systems fail together

The 2024 CrowdStrike-related IT outages showed how a single technical change can become a widespread operational event.[41] AI sharpens that risk because many tools are live services controlled by providers, rather than static products fully owned by financial institutions. A bad update, degraded model, or change in service can hit many users at once.

Coding assistants add a very modern version of the same risk. For example, Google said in April 2026 that 75% of its new code is AI generated[42] – which has risen from about 25% in late 2024. A concern is that if many firms use the same coding tools, they may reproduce

similar code patterns, bugs, security weaknesses or architectural assumptions. Those shared weaknesses could then become common exploit paths – turning software development itself into a hidden channel of contagion.

…Or that shared inputs create shared misreadings

As well as shared models or infrastructure, contagion can also come from shared inputs. If many firms are tracking the same Bloomberg feed, the same credit data, or the same online sentiment signal, a faulty or manipulated input can push multiple systems toward the same shared mistakes. This is especially dangerous where the data are not obviously wrong. Errors (or deliberate fabrications) like duplicated prices, outdated data feeds, or manipulated social media activity can look real enough for AI systems to act on before humans notice them.[43]

And once a false signal enters an AI-enabled workflow, it can travel. For example, a model-generated market summary might feed a risk dashboard – where it shapes alerts or recommendations to review exposures. One model's outputs becomes another's inputs.

Efficiency in normal times – but fragility under stress?

Financial institutions are using AI to improve the way they manage risk: strengthening hedging, detecting anomalies earlier, and making risk management more responsive. But if many firms use similar systems to read the same signals and optimise in similar ways, that improvement can create false reassurance. What looks like better risk management in ordinary times may also be convergence. This concern predates Gen AI, but recent academic literature has added to it.

The nuance, here, is that successful AI may be more systemically important than failed AI. Bad systems get abandoned; good systems spread, gain trust and become embedded. Only later might the common weak point become visible. If those systems are also sitting inside control functions, there is a danger that firms misunderstand their own condition in the same way. And because AI can detect and act on signals almost instantly, crisis time compresses: the window for human intervention can shrink just when it matters most.

Some forms of concentration are more dangerous than others

As the Financial Stability Board's framework[44] for monitoring third-party dependencies suggests, the point is not simply that large providers are risky. A provider can be widely used without being systemically critical – and a less visible service can be critical if many important workflows depend on it and there are few substitutes. As the report frames it, the real vulnerability emerges from the interaction of *criticality, concentration, substitutability,* and *systemic relevance*. The practical question is: if a provider fails, how many important processes stop, how quickly does the disruption spread, and how easily can firms switch to something else.

The mechanisms behind market correlation

The FSB report: "The Financial Stability Implications of Artificial Intelligence"[45] lays out four drivers behind the use of common data and models (the bullets below are quoted verbatim).

- **Herding behaviour:** Market participants imitate data and model choices of others.
- **Network externalities:** The performance of models trained by third parties may improve from interactions with a wider range of end-users and thus incentivise multiple agents to use specific third-party models.
- **Limited choice:** Few data sources and models meet acceptable performance levels. Limited choice could be driven by service provider concentration.
- **Lack of transparency:** Model providers may not disclose their training data sources. End-users could thus unknowingly rely on the same data sources as other financial market participants.

Gen AI sharpens this risk because many firms adopt (and adapt) the same pre-trained models, built on similar architectures and data. Even customised systems may produce correlated outputs. If agentic AI starts triggering actions, not just producing insights, similar systems could move from shaping how firms understand risk to shaping how they respond to it.

The uncomfortable tension: *Parallels with the financial crisis*

Six years ago, this report's predecessor noted: "One of the most alarming revelations of the financial crisis was the degree to which global financial markets were interconnected. How AI might exacerbate the risk of contagion is speculation. But there is a plausible case that these technologies will create new kinds of interconnectedness – at the data level, the IT systems level, and the decision-making level".

That feels even more relevant in the age of Gen AI and agentic systems than it did then. The point is not that the next crisis will replicate the last one, but that there may be similar patterns. Before 2008, many financial institutions thought they were diversified, hedged and sophisticated. In fact they were exposed to the same assumptions about credit risk, liquidity, ratings, and market correlations – and, ultimately, to the same confidence that those risks would not crystallise together. Today, firms may look independent at the surface while relying on the same cloud providers, model families, data feeds, coding tools, and infrastructure underneath.

In 2024, Gary Gensler, then the chair of the US Securities and Exchange Commission, told *Politico:* "I would be quite surprised if in the next 10 or 20 years a financial crisis happens and there wasn't somewhere in the mix some overreliance on one single data set or single base model somewhere".[46]

This is not a fringe concern

The precise systemic risks from AI remain uncertain – which is hardly surprising, given the pace and novelty of the technology. Even so, the subject is high on the macroprudential agenda. In April 2025, the Bank of England's Financial Policy Committee identified four AI-related channels of potential financial stability risk[47]: core decision-making by banks and insurers; use of AI in financial markets; operational reliance on AI service providers; and the changing cyber threat environment.

The FPC analysis made an explicit comparison to the financial crisis. It said: "In the event that large numbers of firms rely on the same open-source model components or data libraries, a significant unknown error or bias could cause many firms to misestimate certain risks and so misprice and misallocate credit as a result. The eventual crystallisation of such a weakness could generate losses for a number of systemic firms, leading to a tightening of credit supply to the real economy, or broader financial contagion through a loss of confidence. This type of scenario was seen in the 2008 Global Financial Crisis, where a debt bubble was partly fuelled by the collective mispricing of risk".

There are broader concerns, here, that firms' use of AI in response to a shock could produce actions that make sense at firm level, but are not based on sufficient information (or the right incentives) to take account of system level outcomes. In autonomous trading, this could go further: models might learn that stress creates profit opportunities and take actions that amplify, or even help trigger, such events.

In April 2026, the FPC suggested that financial system participants have not yet adopted more advanced forms of AI in a manner that would present systemic risk. But it added: "However, risks are likely to increase, potentially rapidly, amid growing intent among financial firms to expand their deployment of advanced AI".

What could happen? A July 2025 paper by the Systemic Risk Centre at the LSE lays out one potential route by which AI could exacerbate a financial crisis.[48] It describes how malicious uses of AI – from market manipulation to attacks on financial infrastructure – could provide the initial shock. Even a small, coordinated misinformation shock can push collective beliefs beyond a tipping point – and trigger a self-fulfilling crisis.

This is linked to AI "wrong-way risk", where systems earn trust by performing well in routine conditions, encouraging firms to use them in more consequential settings. But those same systems may be least reliable when the environment changes sharply, and historical patterns become less useful. This weakness can be compounded by synchronised behaviour: the optimal move for one market participant gives others an incentive to do the same. And the final ingredient is speed. When AI acts as a crisis amplifier, what once unfolded over days or weeks could happen in minutes or hours.

> *"Our analysis suggests that AI will likely lower day-to-day volatility while increasing tail risk — smoothing out short-term fluctuations at the expense of more extreme events. When faced with minor disturbances, AI can absorb shocks and stabilise markets. However, during genuine stress events, the same capabilities that dampen small fluctuations may amplify extreme movements, making crises faster and more intense than those we have experienced previously."*
>
> "Artificial intelligence and financial crises", Systemic Risk Centre, London School of Economics (2025)

The risk framework in this report focuses on risks arising from the use of AI within financial services: how AI affects customers and society; how it reshapes firms' operating environment; and how risks might scale across the financial system.

At least three very significant wider risks sit just outside that framework: AI's impact on energy consumption and the environment, its impact on human work and jobs, and the potential for AI-related asset bubbles. These are broader risks facing society that the sector is exposed to, helps finance, or may amplify through its own activities. This report provides just a brief overview of them – not because they are any less important, but because each is substantial enough to warrant detailed analysis in its own right.

Environmental impact

AI compute demands raise energy and infrastructure concerns.

A fast-rising societal concern around the proliferation of AI is its hunger for energy. Gen AI in particular depends on compute-intensive data centres, implying growing electricity demand, emissions, and water use. The International Energy Agency (IEA) is perhaps the most widely cited source on this. It projects that electricity demand from data centres worldwide is set to more than double by 2030,[49] to around 945 TWh (which is slightly more than the entire electricity consumption of Japan today). The IEA says that AI will be the most significant driver of this increase, with electricity demand from AI-optimised data centres projected to more than quadruple by 2030. The issue is not only emissions: the IEA also warns of grid constraints and rising demand for critical minerals.

The financial services industry's share of AI's environmental footprint is difficult to isolate. Research tends to measure data centre and AI infrastructure impacts overall, not by end-user sector. But the financial sector is certainly not a bystander in this conversation. It is a major user of AI-enabled services; it finances and invests in AI infrastructure; and it is an insurer of climate and infrastructure risk. High on the list of concerns are:

- **Environmental commitments:** Rising AI use could increase energy consumption and emissions in ways that undermine firms' net-zero and sustainability commitments. This is especially sensitive for a sector already under pressure to align financing, underwriting, and investment activity with climate goals.
- **Reputational risk:** Firms may face criticism if their AI use appears wasteful, excessive, or poorly linked to social value. The reputational issue is not just how much compute is used, but whether the use case feels worth the environmental cost.
- **Use-case legitimacy:** Some AI applications, such as high-frequency trading[50] or marginal optimisation of advertising/pricing, may be much harder to justify from an environmental perspective than AI for fraud prevention, accessibility, or climate-risk modelling.

And, as this report has emphasised, compute dependence is more than an environmental issue. AI adoption deepens reliance on data centres, cloud providers, chips, grids, and cooling infrastructure, making it an operational resilience and supply-chain concern as well.

This is a subject that deserves focused attention, not least because financial services is on the front line of environmental risk. For example, climate change has consistently been ranked as one of the top threats facing the global insurance industry (especially reinsurance) in the CSFI's biennial 'Banana Skins' surveys for many years – as concerns grow around whether many climate-related risks are even possible to insure. The industry needs a clearer account of the energy use and infrastructure dependencies created by its growing reliance on AI.

Workforce disruption

Job displacement and disruption to skills and career pathways.

One of the most immediate sources of alarm about AI is its potential to replace jobs across the economy, including roles that once seemed relatively protected by high levels of education and training.

In financial services, it is clear that AI is capable of performing many of the tasks that make up today's human roles. This raises two interpretations that go in opposite directions. One is that previous technological advances have always created more jobs than they replaced, and AI will be no different. The second is that AI *is* fundamentally different because, unlike previous technologies, it leaves less (perhaps much less) space higher up the value chain for humans to go. As one data science head at a large financial institution put it to me: "Humans are still essential, but the question is whether you need 10 or 100". This is one of the biggest questions of our times. (And one which this report will not attempt to answer…)

The survey of 78 senior financial services practitioners and observers conducted for this report asked respondents for their view on the question: "I am concerned about the extent to which AI will replace human talent in my sector over the next five years." 47% agreed with the statement (including 16% strongly), while 33% disagreed (6% strongly).

The comments that came back, however, were less about job losses than about how roles could be reshaped. One respondent said: "I have a strong belief that human talent will continue to play an important

role in banking over the coming years. There will be a clear change in the skills required, moving to a more 'soft skills' approach, with AI reducing the need for technical knowledge".

Another said: "I believe that AI will improve efficiency of processes and complement human intelligence in many ways, but I am concerned about a hollowing out of graduate and junior level roles, including administrative and secretariat roles. This could result in a false economy, because over time businesses will struggle to replace mid-level management / technical staff by failing to nurture a pipeline of juniors".

AI asset bubble risk

Inflated valuations from AI hype.

While there are always risks from hype, genuinely transformative technologies can create enormous economic value and still generate asset bubbles along the way.

The Bank of England recently warned[51] that AI-related stocks now account for a much larger share of US equity indices – "close to levels seen at the peak of the dot com bubble" – and that the exposed universe extends to the whole AI stack: cloud service providers, model developers, specialist chip manufacturers, app developers, data centre operators, and companies that specialise in networking, storage, and cooling systems for data centres. It warned: "If the projected scale of debt-financed AI and associated energy infrastructure investment materialises over this decade, financial stability risks are likely to grow".

Briefly, AI asset bubble risk has at least three channels:

- **Overpricing:** The risk that markets price AI too confidently. AI could prove wholly transformative, and yet the wrong assumptions can misdirect capital, detach asset prices from fundamentals, and create instability when the market's expectations are revised. Overpricing does not just hurt investors: it can channel funding toward the wrong firms, technologies, or infrastructure before the real economics of AI are clear.
- **Amplification:** The risk that finance locks uncertain AI expectations into long-lived projects, debt, and balance-sheet exposures. If banks and investors fund data centres, power projects, or AI-linked firms on overly optimistic assumptions, a market correction could leave them exposed to underused assets and credit losses.
- **Exposure:** The risk that financial institutions become too concentrated in the AI boom. Exposure may sit not only in obvious AI equities, but also in broad indices, pension portfolios, private credit, and so on. If expectations reverse, losses could spread through portfolios that appeared more diversified than they really were.

> *"If routine, low-risk tasks are where people actually learn the ropes, how are they going to learn the technical expertise if they don't apply it? A lot of AI discussion focuses on the idea that automation releases people from menial tasks, meaning they can focus their time and energy on more strategic work and softer behavioural competencies and skills. That is true, but in the context of outputs generated by AI, technical expertise is more important than ever. You need human judgment to really understand, evaluate and assess whether an AI-generated output is accurate or not. And humans have historically been the source of poor decision-making – so it is not about having any human in the loop, but what ethical training you provide to them to make the right decisions and judgments."*
>
> Head of research, insurance sector

Concluding thoughts

This report opened with a discussion of what AI in financial services really means, how it is used, and a brief overview of the potential advantages. It then focused in more detail on fleshing out an 'AI risk map for financial services' – a fundamental refresh of an earlier framework, this time for the Gen AI era.

This mapping is far from the only way to frame risks, and it does not claim to be exhaustive. There is also some overlap between risk categories (which is, perhaps, unavoidable however you cut it). But it is a useful way to trace how AI risk moves through the financial services ecosystem: from the outcomes experienced by customers and society, to the environment in which firms deploy AI, to the system-level dynamics through which risks can scale and spread. It illustrates how the nature of each risk has changed with advances in generative AI (and other frontier technologies), and identifies an "uncomfortable tension" at the heart of each one.

Three essential insights run through the analysis:

- **AI risk in financial services is largely about trade-offs,** so the very same features that make AI useful (prediction, personalisation, scale, complexity, autonomy, etc.) also create the risks that make it hard to govern. The pertinent question, then, is not just whether these risks can be mitigated, but how much risk we are willing to live with in exchange for the benefits. That is why "uncomfortable tensions" are a recurring theme throughout the report".
- **Gen AI has fundamentally changed the risk landscape,** not only because it is prone to hallucination (which is perhaps the most obvious risk), but because it makes AI widely accessible, persuasive, easy to use, and embedded in everyday financial workflows. The 2019 CSFI risk framework, while still relevant, was created before these technologies became sophisticated or widespread – which is why it needed an overhaul.
- **Many of the thorniest risks in this report are ecosystem risks** – not only those explicitly labelled 'system' level, but any risk that can scale, spread, or reinforce other risks across the financial system. They often involve choices that look sensible for individual firms, but can create shared dependencies and fragility across the system. That is also true of risks that may first appear more localised. Over-optimised AI can produce exclusionary outcomes that weaken the social value of financial services. Opaque AI can damage trust in the financial ecosystem as a whole. Misleading outputs can propagate until a seemingly isolated failure becomes a wider source of harm.

The AI economic transition

The financial services sector will not just use AI – it will need to fund, insure and absorb its shocks

The framework introduced in this report has an overarching section at its end: 'The AI Economic Transition'. The key point, here, is that AI risk in financial services is not only about the *use* of AI by financial institutions and other market participants. It is also about the transformation of the wider economy that the industry funds, insures, prices and intermediates. This is critical.

If, as expected, AI transforms how firms produce value, how workers earn income, how infrastructure is built, how assets are priced, how risk is distributed, and many more such things, those changes will flow back into the business – and, indeed, the purpose – of financial services.

There is a striking parallel with another vital challenge the world faces: climate transition risk. In both cases, the issue is not only the direct impact of the underlying force, but the economic reallocation it triggers. AI creates new sources of growth and productivity, but it may also shift value between firms, sectors, workers, asset owners and countries. The financial system will be both an enabler of that transition and exposed to its consequences.

The potential channels are many, and practically impossible to bound. Still, they are likely to include:

- **Labour income:** AI's effects on employment security and wage growth will flow through mortgages, savings, pensions, credit risk and consumer vulnerability.
- **Sectoral reallocation:** some firms and sectors may benefit from AI, while others face disruption or obsolescence – changing credit quality, investment returns and demand for financial services.
- **Returns to capital:** if AI shifts more income from workers to the owners of companies and assets, it could deepen inequality and turn access to financial ownership into an even sharper dividing line.
- **Asset valuations:** markets may overprice some parts of the AI economy, while underestimating where value ultimately accrues.
- **Insurance and risk transfer:** AI may create new insurable risks, while making some losses harder to attribute, price, or pool.
- **Public finances:** changes to employment, productivity, and tax bases could alter fiscal positions and sovereign-risk assumptions.

There may be benefits to taking a leaf from climate scenario analysis approaches already in use (without overplaying the analogy). The point would not be to predict a single AI future, but to test how institutions, portfolios and business models might behave under materially different scenarios. AI does not have an equivalent of emissions pathways or temperature outcomes. But the approach could be broadly similar: identify the assumptions already embedded in markets and business models, test how those assumptions could fail, and ask where risks would concentrate under different pathways.

Scenario analysis is already becoming part of the supervisory response. In April 2026, in its response[52] to the House of Commons Treasury Committee's 2026 report on artificial intelligence in financial services, the Bank of England said it is using scenario analysis to test plausible macroeconomic and financial-market outcomes from AI investment, development and adoption. It is also incorporating AI into cyber and operational testing, and examining whether AI trading agents could herd in ways that amplify market stress.

In all of this, it is important to remember that risk does not lie only in the characteristics of the technology, but in its use and implementation.

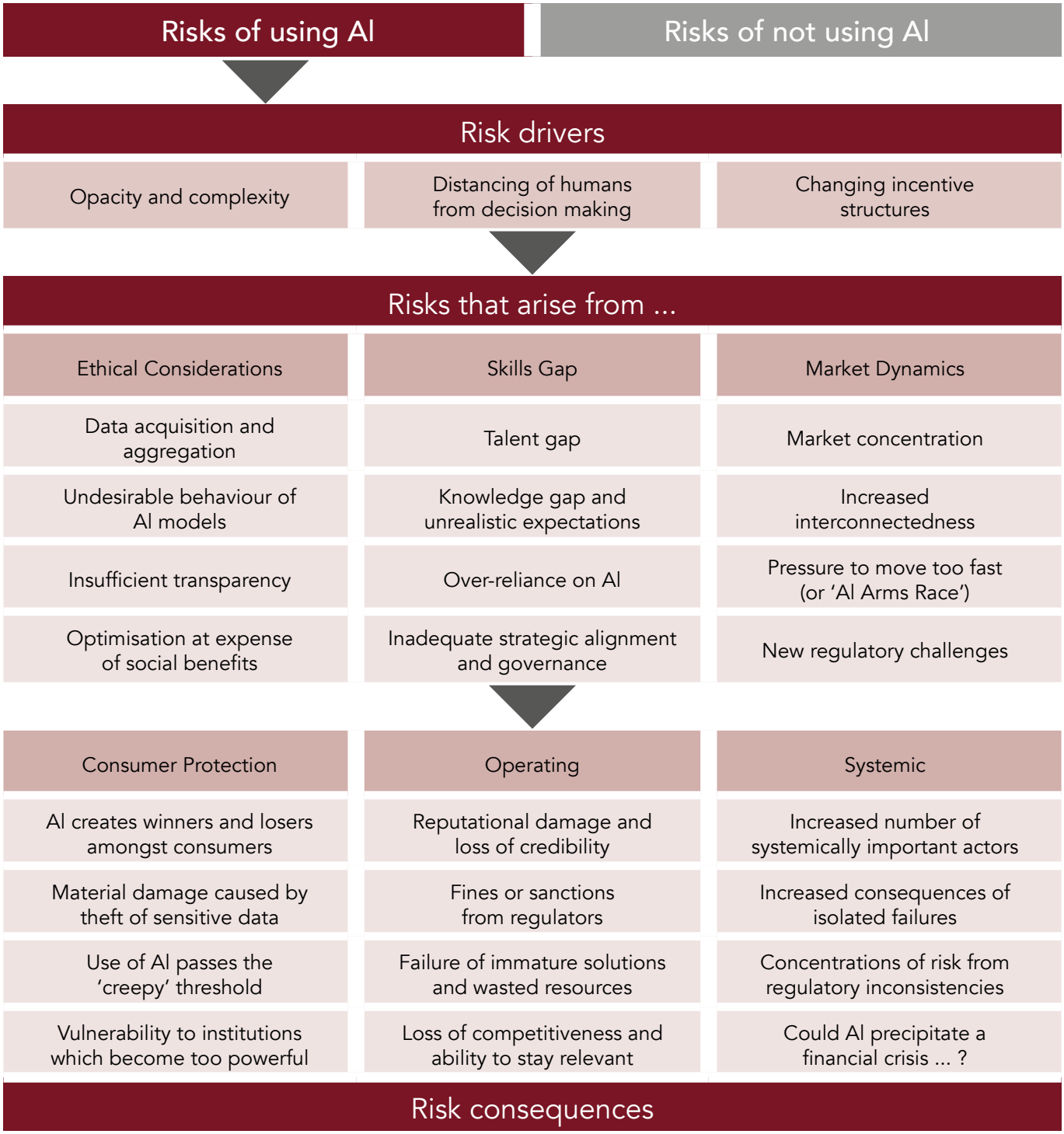
The 2019 predecessor to this report concluded that AI outcomes ultimately depend upon humans, not machines. Much of this report has emphasised how the machinery of AI has changed, particularly since the arrival of Gen AI: how it has become more sophisticated and, in many respects, more unpredictable. But much of the change is also about the ubiquity of use and the way AI's fluent outputs can create, rightly or wrongly, an *impression* of sophistication.

The future of AI is impossible to predict, and serious conversations about artificial general intelligence now seem disquietingly plausible. But, at least for now, the risk is as much about humans relying on AI too much, or in the wrong ways.

Appendix: the CSFI's 2019 risk framework

The CSFI's 2019 report, *It's Not Magic: Weighing the Risks of Financial Services*, introduced a framework of 12 key risks in three categories: the ethical challenges that can arise; the risks around the implementation of these technologies; and how widespread adoption of AI might affect market dynamics. It looked at characteristics of AI technologies that help explain why AI is different from other forms of automation – the "risk drivers" – and how these risks might affect consumers, institutions, and the financial system as a whole – the "risk consequences".

Chart 1: Risks to the financial services industry arising from AI



References

1. "**Did a Bug in Deep Blue Lead to Kasparov's Defeat?**" *CNET.* https://www.cnet.com/culture/did-a-bug-in-deep-blue-lead-to-kasparovs-defeat/
2. "**'A Feedback Loop with No Brake': How an AI Doomsday Report Shook US Markets.**" *The Guardian.* https://www.theguardian.com/technology/2026/feb/24/feedback-loop-no-brake-how-ai-doomsday-report-rattled-markets
3. See, for example, **"Managing explanations: how regulators can address AI explainability,** Financial Stability Institute" (September 2025) https://www.bis.org/fsi/fsipapers24.pdf The report notes: "With the advent of newer-generation reasoning models, the focus has shifted. These newer models aim to reproduce human-like reasoning patterns more directly. Importantly, while the explanations generated by these models may appear to mimic how humans reason, they do not necessarily reflect how the models actually arrive at their conclusions. In other words, producing an appealing explanation does not guarantee that it represents the model's internal reasoning process."
4. For an in-depth analysis, see, for example, **"LLMs Will Always Hallucinate, and We Need to Live With This"**, *Banerjee, Agarwal, Singla, United We Care* https://arxiv.org/pdf/2409.05746
5. See, for example, **"AI Safety and Automation Bias"**, *Centre for Security and Emerging Technology (November 2024)* https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/
6. This has also been highlighted by regulators. For example, see: **"Consultation Paper of Guidelines on Artificial Intelligence Risk Management"**, *Monetary Authority of Singapore (November 2025)*. https://www.mas.gov.sg/publications/consultations/2025/consultation-paper-on-guidelines-on-artificial-intelligence-risk-management
7. "**Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector**", *Financial Stability Board (October 2025).* https://www.fsb.org/uploads/P101025.pdf The five layers are (the following is a truncated quote from the report): "1) Hardware: Computing chips necessary for training and using many AI models, including graphics processing units (GPUs) and other AI-specialised chips; 2) Computing infrastructure: Primarily revolves around cloud services, which are crucial for the development and distribution of state-of-the-art GenAI applications. 3) Training data: Large datasets require significant resources to aggregate and manage. While some data is publicly available, much of it is proprietary or sourced from third party aggregator 4) Pre-trained foundation models: Large-scale AI models, such as LLMs, that are trained and disseminated by third parties, such as AI firms 5) User-facing applications: The layer that governs how end-users interact with AI models for specific use cases."
8. See, for example, **"The agentic AI landscape and its conceptual foundations"**, *OECD.* https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf
9. "**Agentic AI and consumers**", *Competition and Markets Authority (2026)* https://www.gov.uk/government/publications/agentic-ai-and-consumers/agentic-ai-and-consumers
10. AGI is much discussed, although its meaning is notoriously difficult to pin down precisely. Wikipedia defines it as "a hypothetical type of artificial intelligence that matches or surpasses human capabilities across virtually all cognitive tasks". That gives the general flavour, but the definitional rabbit hole is deep. More technical treatments try to distinguish different 'levels' of AGI by looking at breadth, performance, autonomy and other dimensions of capability.
11. "**Artificial intelligence in UK financial services – 2024**", *Bank of England and Financial Conduct Authority* https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024
12. "**The 2026 Global AI in Financial Services Report: Adoption, impact and risks**", *Cambridge Centre for Alternative Finance, University of Cambridge* https://www.jbs.cam.ac.uk/wp-content/uploads/2026/04/ccaf-2026-04-28-global-ai-in-financial-services-report.pdf
13. Respondents included 203 fintechs, 149 traditional financial institutions, 146 AI vendors and 130 regulators
14. See, for example, **"Artificial Intelligence Consortium minutes" – February 2026,** *Bank of England*. https://www.bankofengland.co.uk/minutes/2026/february/ai-consortium-minutes-9-february-2026 The minutes from the consortium say: "Members discussed the relationship between risk and return in financial services use cases. It was suggested that current deployments tend to focus on lower-risk, lower-return applications, but that firms may, over time, move towards higher-risk use cases with greater potential returns".
15. See, for example, **"Using Artificial Intelligence methods to predict financial market stress"**, *Systemic Risk Centre (January 2026)* https://www.systemicrisk.ac.uk/using-artificial-intelligence-methods-predict-financial-market-stress
16. The **Pensions and Lifetime Savings Association*** writes: "Savers may… turn to AI tools already available to the mass market – like ChatGPT or Gemini – in helping with financial decision making. This in itself poses serious risks to consumers. Models like these have not been developed specifically for this purpose and are far more likely to produce incorrect output. In many cases, their training data includes a large amount of text scraped from the internet, including websites such as Wikipedia. Where there are falsities and errors in the data that is inputted to the model, errors are likely to be produced in the output." https://committees.parliament.uk/writtenevidence/140271/default/
And Hymans Robertson LLP*, the pension services firm, writes: "Members are already turning to public AI tools like ChatGPT to interpret pension information—tools lacking appropriate regulatory guardrails". https://committees.parliament.uk/writtenevidence/139792/html/
17. For example, Lloyd's Market Association* writes: "Where BigTech firms have used partnerships with incumbent insurers to bring about new products and services, they have also gained long-term access to product performance and claims data which builds their understanding and provides the 'quality data' they need to enter the market. This could enable BigTech firms to leverage their market position to enable advanced risk modelling and use third-party profiling data. While granular data can lead to more individual pricing, it may reduce traditional risk pooling, potentially excluding vulnerable customers. This may initially increase competition but could later lead to reduced consumer choice and higher margins in the long term." https://committees.parliament.uk/writtenevidence/140313/html/
18. For example, the **Centre for the Public Understanding of Finance (PUFin),** The Open University*, writes: "…However, there is the risk of adverse outcomes for two other groups. The first group is those for whom additional AI-generated insights cause their risk rating to deteriorate. The second group is consumers who for whatever reason (including digital exclusion) do not or cannot participate in giving access to the data used to feed the AI tools; they may, by default

be deemed higher risk because of the lack of data (a new generation of 'thin' files)." https://committees.parliament.uk/writtenevidence/140297/html/

19. For example, the Charted Insurance Institute* writes: "There is the potential for a two-tier system to emerge where AI assistants help to deliver consumer benefits through productivity gains for professionals within well-managed boundaries, whereas other consumers are left to interact with 'raw' AI and develop their own mitigations to limitations or errors." https://committees.parliament.uk/writtenevidence/140567/default/
20. The Centre for Protecting Women Online* writes: "Paradoxically, one safeguard against bias is the inclusion of protected attributes (e.g. gender, ethnicity) in the modelling process—not for use in decision-making, but to monitor and mitigate potential bias. Most current bias detection and mitigation methods require access to such sensitive information to evaluate and correct disparities." https://committees.parliament.uk/writtenevidence/140251/default/
21. For example, the Working Group on fAIr Credit, Credit Research Centre, University of Edinburgh*, writes: "It is currently unclear what constitutes a fair or unbiased outcome. Multiple definitions of bias exist in the technical literature (Kozodoi et al, 2022), yet none are codified in regulation. Clarifying which fairness metrics and benchmarks are appropriate in specific contexts would assist decision-makers in navigating these complexities. A related issue is the lack of quantification of proportionality in the legal definition of indirect discrimination. Many technical definitions of fairness are mutually incompatible, and it is not possible to satisfy all simultaneously. Despite this, fairness is often referred to in abstract terms, without identifying which specific definitions or principles are being invoked. There is limited understanding of how technical measures of fairness align with public perceptions."
22. An illustrative example is given by Dr Alison Lui, Dr Lola Durodola and Dr George Lamb*, who write: There are three main categories of people which require explainability in credit scoring. First, loan officers that are said to prefer local sample-based explanations. This involves comparing the unsuccessful applicant's profile against other similar profiles. Second, rejected loan applicants that are said to prefer local feature-based explanations. This involves providing individual case specific reasons. Third, regulators or data scientists that are said to prefer global model explanations. This is because by having a global picture of logic and reasoning used by the model, it enables regulators to ensure that the model is fair and consistent in making decisions. https://committees.parliament.uk/writtenevidence/138946/pdf/
23. "Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models." *Smith, A., Greaves, F. and Panch, T.* https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000388
24. For example, Ms Ana Isabel Canhoto, Professor of Digital Business, University of Sussex*, writes: "Financial service organisations' perspective is limited to the transactions they process. While some customers may use a single financial institution for all their banking, most use multiple providers, meaning each organisation only sees a part of the customer's financial activity. Since these organisations do not share customer data due to legal and strategic reasons, their datasets will always be incomplete. Consequently, they may fail to recognise the significance of certain transactions or misinterpret others." https://committees.parliament.uk/writtenevidence/138936/html/
25. For example, the Institute and Faculty of Actuaries* writes: "GenAI hallucination could reduce the benefit of any AI application where a result that is 'largely correct' is not good enough - this is certainly the case in a financial services context. Apparent efficiencies in using AI could be offset where GenAI outputs are having to be interrogated - with workarounds necessary - to counteract hallucinatory impacts." https://committees.parliament.uk/writtenevidence/140597/html/
26. For example, see: **"Prompt Injection Is Not SQL Injection (It May Be Worse)."** ***National Cyber Security Centre. Chismon, Dave. (December 2025)***. https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection
The article argues that prompt injections are a dangerous new class of risk compared to the more familiar problem of 'SQL injection' – and "may never be totally mitigated in the way SQL injection attacks can be."
27. See, for example, **"The Deepfake Blindspot in AI Governance"**, ***LSE International Development, Ntow, Rachel, Dec 2025***. https://blogs.lse.ac.uk/internationaldevelopment/2025/12/04/the-deepfake-blindspot-in-ai-governance/
28. See, for example, **"The Mosaic Effect: Why AI Is Breaking Enterprise Access Control"**, ***Security Boulevard*** https://securityboulevard.com/2025/11/the-mosaic-effect-why-ai-is-breaking-enterprise-access-control/
29. This risk is explicitly referenced in a **"Public Statement on the use of AI in the provision of retail investment services"**, ***European Securities and Markets Authority (ESMA).*** https://www.esma.europa.eu/sites/default/files/2024-05/ESMA35-335435667-5924__Public_Statement_on_AI_and_investment_services.pdf The statement says: "It is important to note the Statement would aim not only to address scenarios where AI tools are specifically developed or officially adopted by the investment firm or bank but also extends to situations involving the use by firm staff of third-party AI technologies (such as Chat GPT, Google Bard, and others) with or without the direct knowledge and approval of senior management."
30. "Artificial Intelligence and Economic and Financial Policymaking: A High-Level Panel of Experts' Report to the G7" *(December 2024)* https://www.dt.mef.gov.it/export/sites/sitodt/modules/documenti_it/HLPE-Report-on-AI.pdf
31. "The AI Public-Private Forum: Final report", *Bank of England (February 2022)* https://www.bankofengland.co.uk/research/fintech/ai-public-private-forum
32. See, for example,: **"Generative AI a stumbling block in EU legislation talks -sources"**, *Reuters (December 2023)* https://www.reuters.com/technology/generative-ai-stumbling-block-eu-legislation-talks-sources-2023-12-01
33. "Treasury Committee: Oral evidence: AI in financial services, HC 684", *House of Commons (October 2025)* https://committees.parliament.uk/oralevidence/16748/html/
34. "FCA announces second cohort for AI Live Testing", *FCA (April 2026)* https://www.fca.org.uk/news/press-releases/fca-announces-second-cohort-ai-live-testing
35. "Council agrees position to streamline rules on Artificial Intelligence", *Council of the EU (March 2026)* https://www.consilium.europa.eu/en/press/press-releases/2026/03/13/council-agrees-position-to-streamline-rules-on-artificial-intelligence
36. "Supervisory Guidance on Model Risk Management", *Federal Reserve, FDIC, OCC (April 2026)* https://www.federalreserve.gov/supervisionreg/srletters/SR2602a1.pdf
37. "The New Physics of Financial Services", *World Economic Forum (August 2018)* https://www3.weforum.org/docs/WEF_New_Physics_of_Financial_Services.pdf
38. "Growth placed at the heart of regulators' remit alongside new measures to boost scrutiny and transparency", *Department for Business and Trade (October 2025)* https://www.gov.uk/government/news/growth-placed-at-the-heart-of-regulators-remit-alongside-new-measures-to-boost-scrutiny-and-transparency
39. "Financial Stability in Focus: Artificial intelligence in the financial system", *Bank of England (April 2025)*
The BoE's Financial Policy Committee says: "Another source

of risk is the potential for such models to facilitate collusion or other forms of market manipulation. Given the ability of some AI models to learn dynamically in multi-agent environments, and challenges around the explainability of model outputs, such adverse behaviours might emerge without the human manager's intention or awareness." https://www.bankofengland.co.uk/financial-stability-in-focus/2025/april-2025

40. For example, see **"The AI Supply Chain"**, *Leonardo Gambacorta and Vatsala Shreeti, Bank of International Settlements (March 2025)*. The report notes: The AI supply chain consists of five key layers: hardware, cloud infrastructure, training data, foundation models and AI applications. At the moment, the market structure of the first two of these layers exhibit significant concentration. The market for end-user facing AI applications is flourishing with the availability of many new applications across sectors (competition for the market) but winner takes all dynamics can easily emerge, as in the case of other digital platforms https://www.bis.org/publ/bppdf/bispap154.pdf
41. See, for example, **"At Least 750 US Hospitals Faced Disruptions During Last Year's CrowdStrike Outage, Study Finds"**, *Wired (June 2025)* https://www.wired.com/story/at-least-750-us-hospitals-faced-disruptions-during-last-years-crowdstrike-outage-study-finds/
42. **"Google says 75% of the company's new code is AI-generated"**, *Business Insider (April 2026)* https://www.businessinsider.com/google-ai-generated-code-75-gemini-agents-software-2026-4
43. See, for example, **"Global Financial Stability Report"**, *IMF (2024)* https://www.imf.org/-/media/files/publications/gfsr/2024/october/english/textrevised.pdf
44. **"Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector"**, *FSB (October 2025)* https://www.fsb.org/uploads/P101025.pdf
45. **"The Financial Stability Implications of Artificial Intelligence"**, *FSB (November 2024)* https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/
46. **"Gensler's warning: Unchecked AI could spark future financial meltdown"**, *Politico (March 2024)* https://www.politico.com/news/2024/03/19/sec-gensler-artificial-intelligence-00147665
47. **"Financial Stability in Focus: Artificial intelligence in the financial system"**, *Financial Policy Committee (April 2025)* https://www.bankofengland.co.uk/financial-stability-in-focus/2025/april-2025
48. **"Artificial intelligence and financial crises"**, *Jon Danielsson and Andreas Uthemann, Systemic Risk Centre, London School of Economics (July 2025)* https://researchonline.lse.ac.uk/id/eprint/128657/1/x.pdf
49. **"Energy demand from AI"**, IEA https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
50. See, for example: **"Environmental Implications Of AI-Enabled Algorithmic Trading: A Sustainability Perspective from Emerging Markets (India)"**, *Harsh Raghuwanshi and Dr Roli Pradhan, International Journal of Environmental Sciences (2025)* https://theaspd.com/index.php/ijes/article/view/8723/6302
51. **"All chips in! Would a fall in AI-related asset valuations have financial stability consequences?"**, *Bank of England (October 2025)* https://www.bankofengland.co.uk/bank-overground/2025/all-chips-in-ai-related-asset-valuations-financial-stability-consequences
52. **"AI in financial services: Responses to the Committee's Fifteenth Report"**, *House of Commons Treasury Committee* https://committees.parliament.uk/publications/52647/documents/292955/default/

About the author

Keyur Patel is a research associate of The London Foundation for Banking & Finance, and has examined AI risks since the mid-2010s, working with research organisations and practitioners. More broadly, he has authored/co-authored more than a dozen CSFI "Banana Skins" reports on the main risks facing insurance, financial inclusion, and banking, as well as the 2019 predecessor to this report: "It's Not Magic: Weighing the Risks of AI in Financial Services".

The London **Foundation** for Banking & Finance | CSFI Centre for the Study of Financial Innovation

Charity number: 297107
Royal Charter number: RC000719

A charity incorporated by Royal Charter

We are **The London Foundation for Banking & Finance**, a charity incorporated by Royal Charter dedicated to supporting knowledge and lifelong education in financial services.

Contact us: enquiries@lfbf.org.uk
Discover more: www.lfbf.org.uk
Follow us: The London Foundation for Banking & Finance

05/26