Evaluating Emerging AI Systems in Financial Services: Insights from the Turing–FCA Workshop

Naman Goel
The Alan Turing Institute and University of Oxford
naman.goel@alumni.epfl.ch https://goelnaman.github.io/

The finance sector, which comprises financial services and financial institutions, is designated as one of the UK's 13 Critical National Infrastructure (CNI) sectors.[1] The resilience of this infrastructure is a central priority in the UK Government Resilience Action Plan.[2] At the same time, financial services are among the fastest adopters of artificial intelligence (AI) technologies.[3] The sector has extensive experience in risk management and operational resilience, providing a strong foundation for the evaluation and governance of emerging AI systems. However, the rapid evolution of AI—particularly the emergence of more autonomous, agentic systems—also raises new questions for this critical sector.[4]

On 17th April 2026, researchers from the Alan Turing Institute convened a workshop on *Evaluating Emerging AI Systems in Financial Services*. The aim of the workshop was to understand how advanced AI is being deployed across financial institutions, what are the emerging risks, what challenges practitioners face in evaluating these systems, where current methods fall short and how these new challenges can be addressed. It brought together perspectives from academia, industry, and regulators, moving beyond abstract principles toward practical approaches to testing and assurance.

The workshop was organised in collaboration with the Financial Conduct Authority (FCA). The organising team included Naman Goel (Turing/Oxford), Carsten Maple (Turing/Warwick), Henrike Mueller (FCA), Lukasz Szpruch (Turing/Edinburgh) and Tony Zemaitis (Turing). More than 30 leaders from major financial institutions, including Aviva, Barclays, Bank of England, Deutsche Bank, FCA, HSBC, Lloyds Banking Group, Morgan Stanley, NatWest Group, and Starling Bank participated in the workshop.

This document provides the author's synthesis of the main insights generated during the workshop. Any inaccuracies in this document are the responsibility of the author and should not be ascribed to speakers, panellists, participants or any institution. In line with the Chatham House Rule, no remarks in this document have been linked to any specific invited participant.

Academic Presentation

Following opening remarks from the FCA's Henrike Mueller, the workshop began with an academic presentation by Turing's Lukasz Szpruch. Lukasz presented a capability-centric approach to governing agentic AI, focusing on what systems can do rather than specific models or use cases. This was joint

[1] https://www.npsa.gov.uk/about-npsa/critical-national-infrastructure
[2] https://www.gov.uk/government/publications/uk-government-resilience-action-plan/uk-government-resilience-action-plan-html#improving-the-resilience-of-critical-national-infrastructure
[3] https://committees.parliament.uk/publications/51128/documents/283671/default/
[4] https://committees.parliament.uk/publications/52647/documents/292955/default/

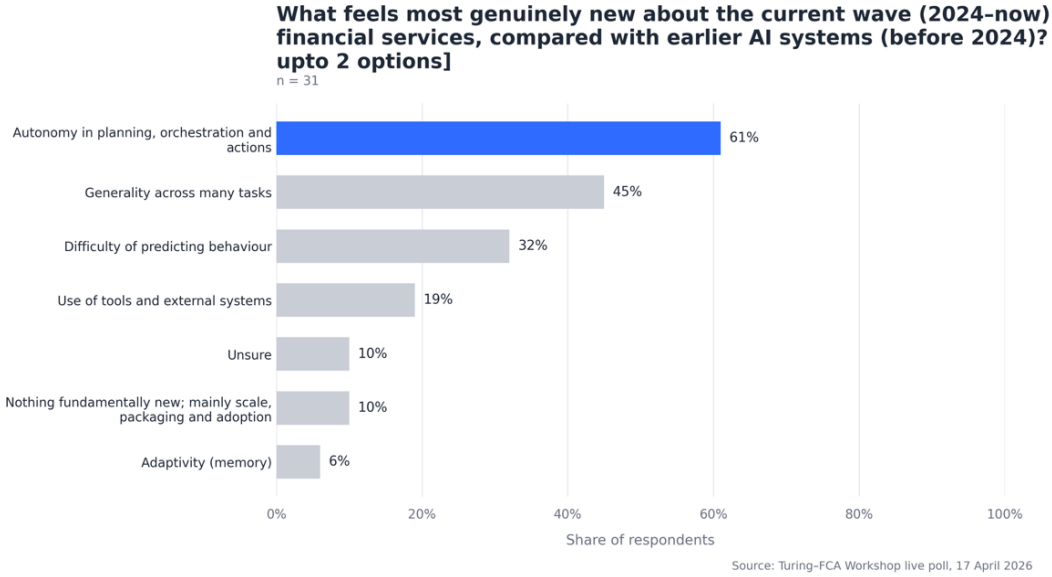
work with Agus Sudjianto, Tanveer Bhatti and Gary Ang, and the paper is available publicly. [5] It introduces a framework that links capabilities to associated risks and controls, offering a more scalable way to manage increasingly autonomous systems. This provides a structured basis for evaluating and governing AI systems whose behaviour extends across multiple tasks and contexts.

The presentation sparked a rich discussion with the audience around how capability-centric governance could be operationalised in practice: what constitutes a "capability" in concrete terms, how such abstractions can be meaningfully defined and standardised and how is this approach different from the currently prevalent approaches. Participants also reflected on the notion of a "risk delta"—how risk evolves when moving from isolated capabilities to full systems or trajectories, and what additional effort is required to account for this in validation.

Live Poll

Turing's Naman Goel conducted an anonymous poll, which participants engaged with actively.[6,7] The poll results help map the current landscape in the industry and indicate where future research efforts should be directed. Main takeaways from the poll are as follows:

1. Autonomy and generality are seen as the main new features of emerging AI systems in financial services.



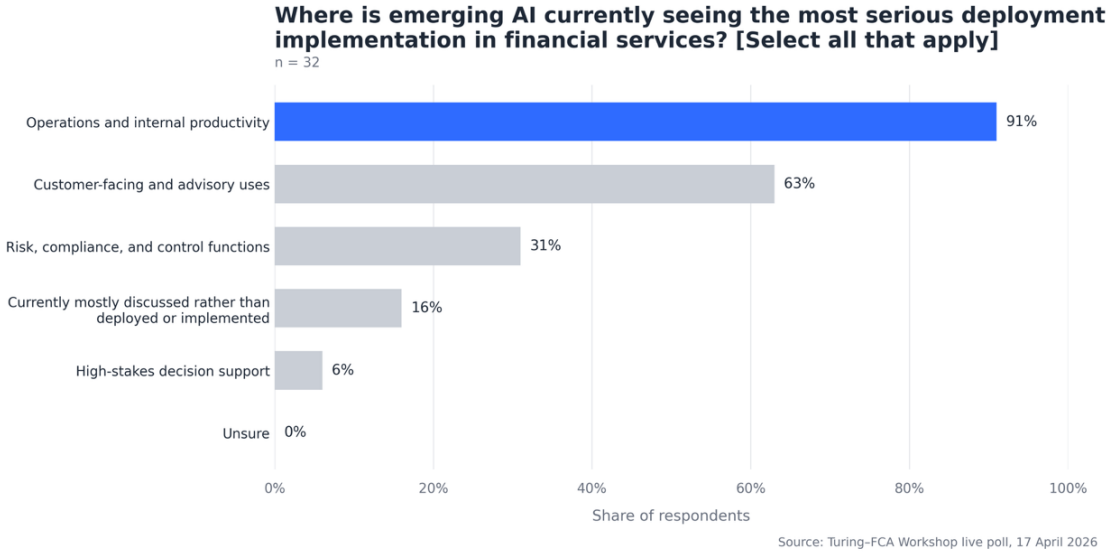
[5] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6567199

[6] The audience represented a balanced mix of industry, regulators, and academia, with a slight skew toward risk/compliance and technical roles. Most participants were directly involved in deploying or governing AI systems, indicating a highly practitioner-heavy audience. See Appendix A for more details.

[7] Participants were asked to answer based on personal observations or assessments, rather than second-hand information where possible.

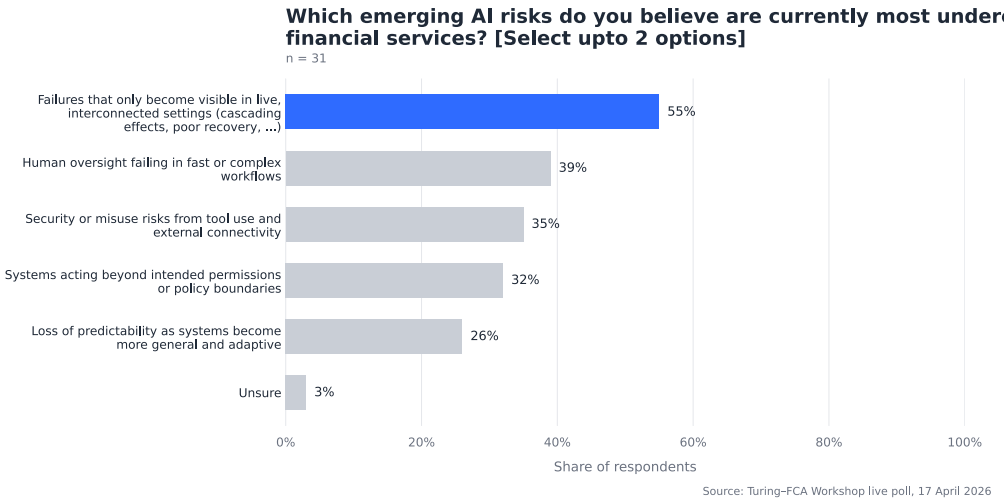
2. Emerging AI systems are most widely deployed in operations and internal productivity, with significant emerging use in customer-facing functions.

A prominent example that was highlighted by some of the participants was software engineering. These observations mirror what has been observed in related work too.[8,9]



3. Participants believed failures that become visible only in live, interconnected environments is the most underestimated risk, with other risk types close behind.

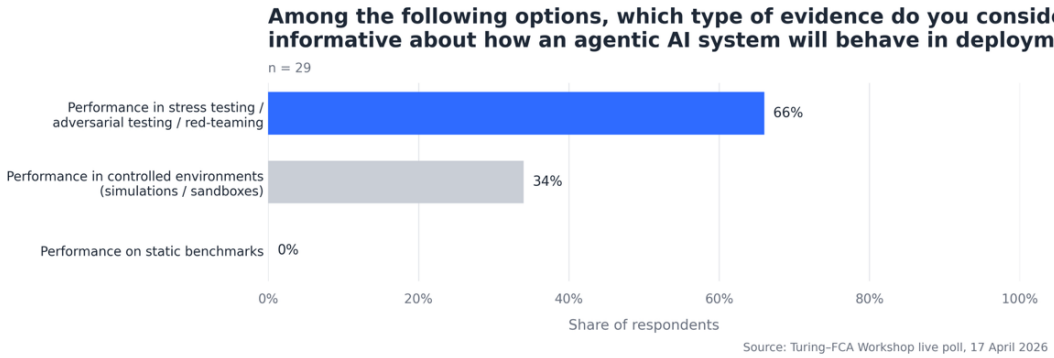
Participants also raised other potentially underestimated risks that were not listed among the options in the poll: cost (e.g. inference cost) of deploying larger models, increasing dependence on technology and potential deskilling of the workforce.



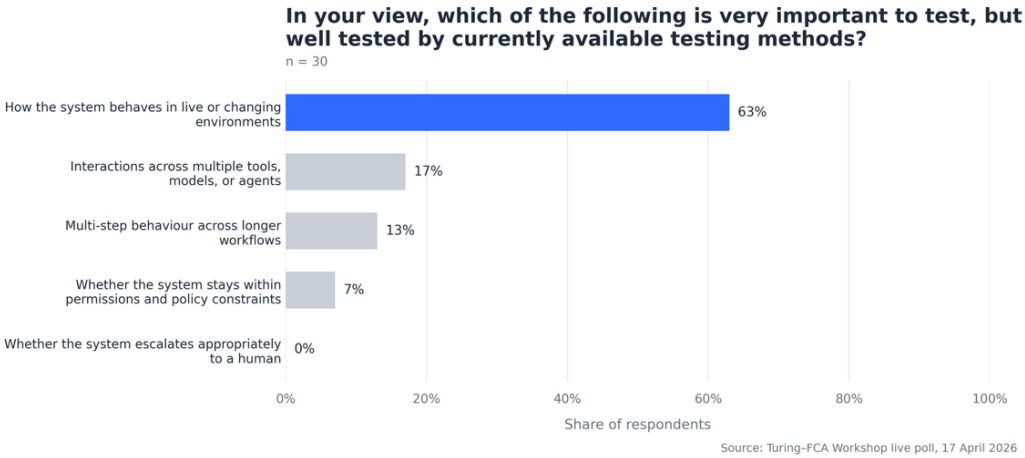
[8] https://arxiv.org/abs/2512.04123
[9] https://arxiv.org/pdf/2603.23802

4. Performance in stress testing and adversarial testing is considered more informative about deployment behaviour of agentic AI than performance on benchmarks or in sandboxed environments.[10]

With other forms of evidence (e.g. live testing) excluded, responses reflect a preference for adversarial evaluation approaches under practical constraints.



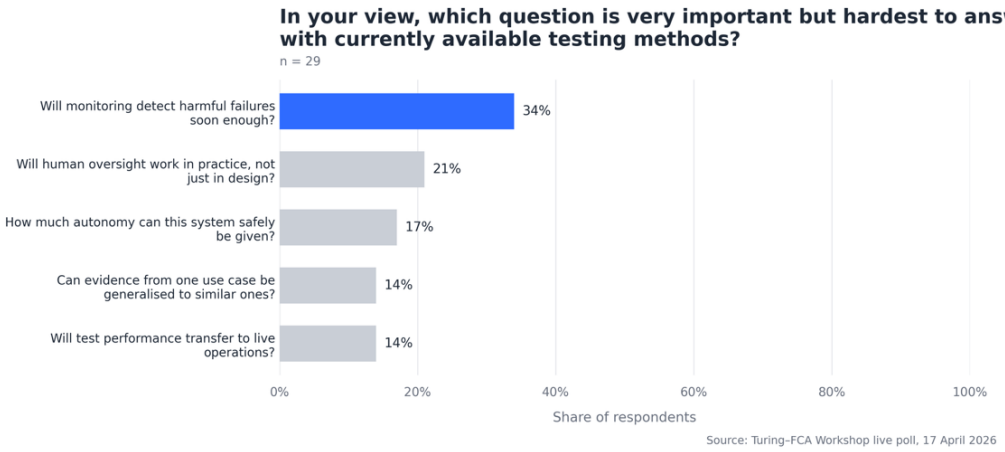
5. System behaviour in live or changing environments is seen as important and hard to test sufficiently with current approaches.



6. Whether monitoring can detect harmful failures in time is one of the hardest questions to answer credibly with current approaches.

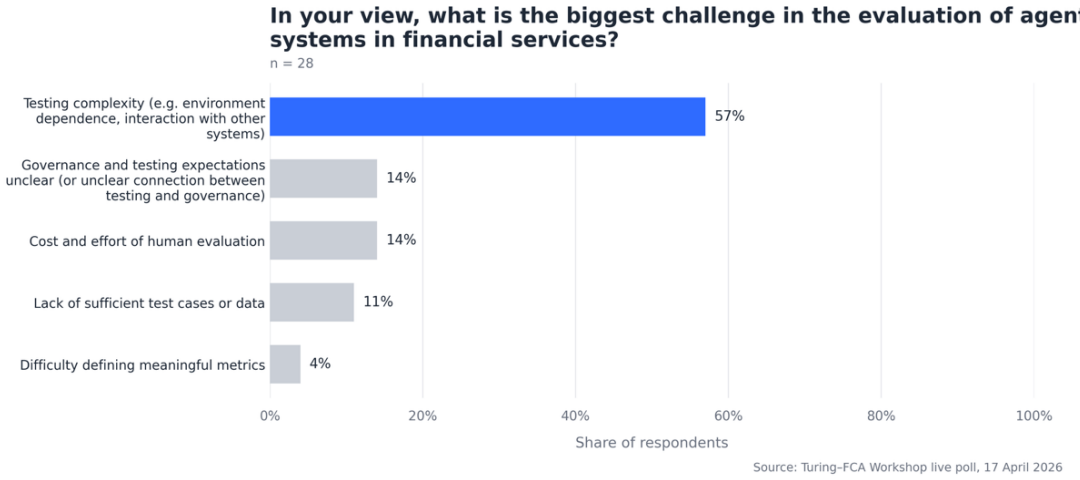
This result may partly reflect the difference between pre-deployment testing approaches (e.g. using batches of labelled test samples) and the methods required to detect subtle failure patterns early during continuous monitoring.

[10] Participants were provided a reference definition of "agentic AI": *LLM based systems that have some or all of the following features: dynamically plan actions for given context, use tools, have memory, execute multi-step workflows with limited human intervention*.

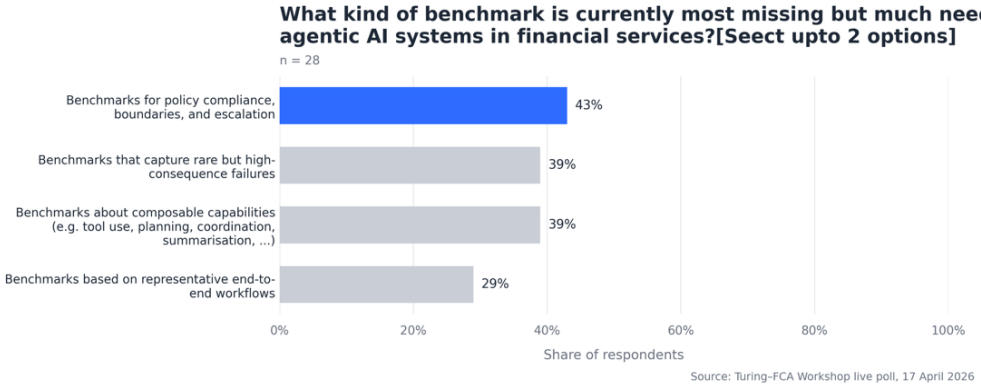


7. Testing complexity, e.g. due to environment dependence and interactions across systems, is among the biggest challenges in evaluation.

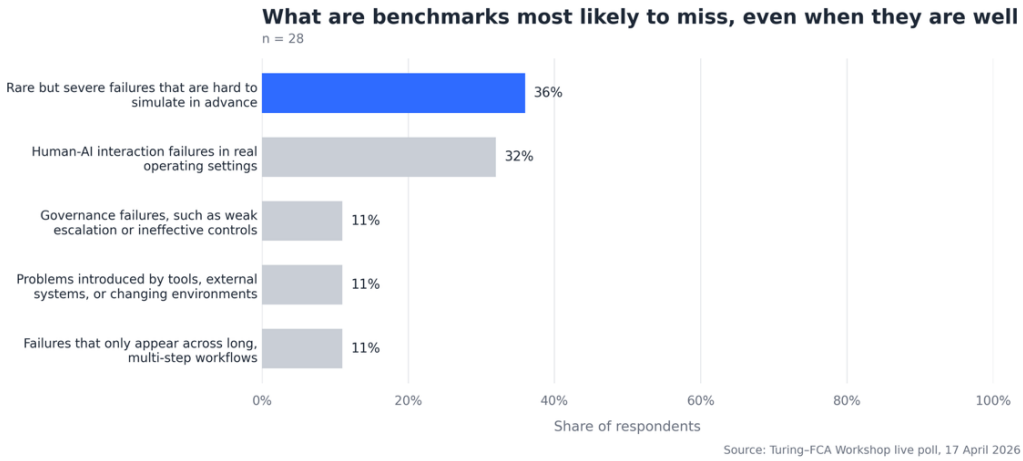
Participants also highlighted the difficulty of evaluating 3rd party systems and components, which is somewhat related to the top option in this poll.



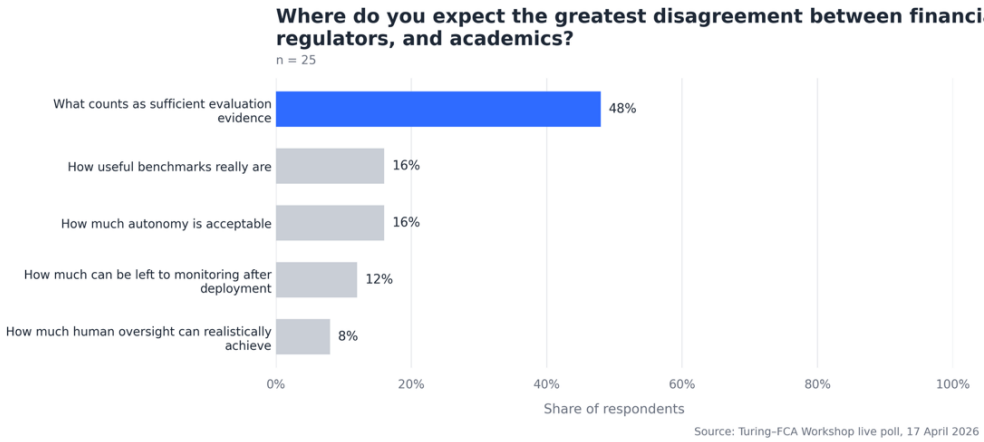
8. Benchmarks for policy compliance, boundaries, and escalation, as well as those capturing rare but high-consequence failures, are seen as missing but much needed.



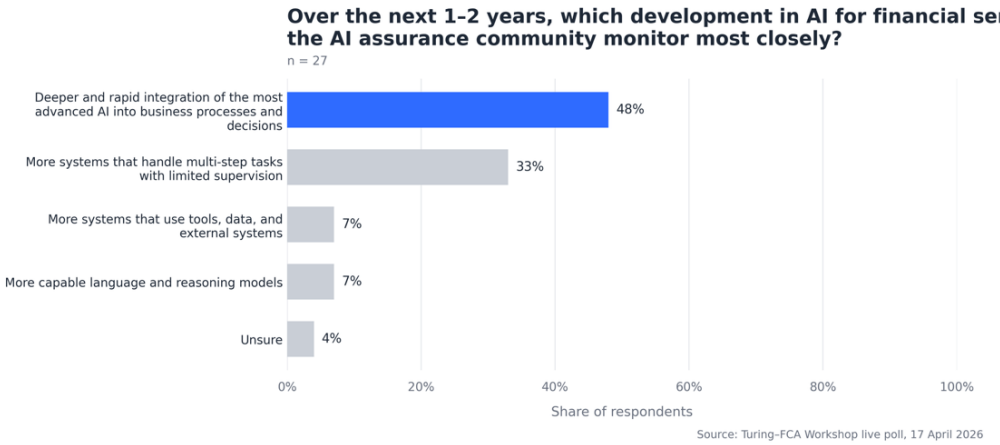
9. Fundamental limitations of benchmarks include missing rare, hard-to-simulate severe failures and real-world human–AI interaction issues.



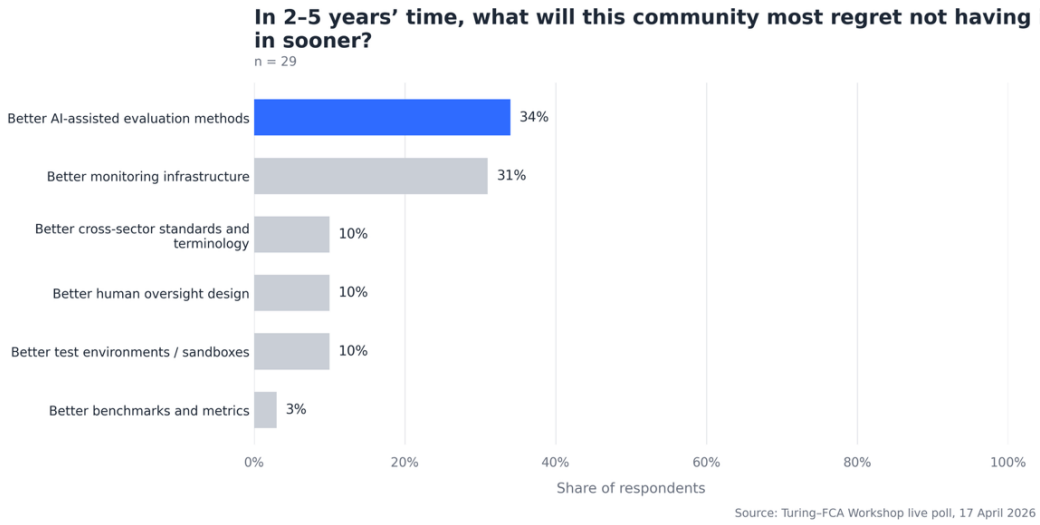
10. The greatest disagreement is expected among financial institutions, regulators, and academics on what counts as sufficient evaluation evidence.



11. The integration of advanced AI into business processes is the key development to monitor.



12. Underinvestment in AI-assisted evaluation methods and monitoring infrastructure is expected to be the biggest regret.



Panel Discussion with Industry Experts

Turing's Carsten Maple chaired a panel discussion bringing together three current and former heads of model risk management from some of the world's largest financial services firms.

The panel discussion highlighted a clear shift in financial services AI: chat and RAG were seen as yesterday's news, with attention now moving to agentic AI, often multi-modal, and expected to handle logical rather than only semantic tasks. At the same time, there was a clear sense of a divide between Silicon Valley and the financial world: technology is advancing quickly, but is often not ready for adoption in regulated settings.

Most current deployments were described as still relatively low risk, with limited autonomy, but with less human supervision than before. Models are also much improved compared to, say 6-18 months ago, shifting validation from identifying obvious failures to detecting more subtle weaknesses. This is happening alongside a strong business desire for customer-facing use-cases.

On new challenges in testing, the discussion noted the cost of validation, especially as systems move from advisory to decision making. Existing principles such as risk tiering still apply, but new challenges also include controls evaluations, deciding how much human in the loop is needed, identifying what will break the system, and testing edge cases. There was particular concern about sensitivity to input, transferability across use-cases and contexts, and the fact that brute-force use-case level testing is not practical for agentic AI. The discussion also noted while testing needs ground truth, human in the loop is not a scalable approach for generating ground truth. Because of cognitive mismatch, people cannot be relied on to generate or review enough test cases. This makes automatic test generation increasingly important. But automated judging is not straightforward: LLM as judge may work for semantic tasks

such as summarisation, but is much less suitable for agentic systems, where one must assess decisions and logic. More structured approaches, including knowledge graph-based approaches, were seen as promising. Human input was seen as most useful for calibration and for architecting the overall approach.

On sufficiency of testing, the discussion stressed that it depends on the use-case worst-case scenario: in some settings, even 5% errors can be huge. This led to calls for tail testing, active learning, interactive testing, sequential design, and explicit thresholds for metrics such as sensitivity to prompts, consistency, and transparency. Runtime governance and harness design also needs to be tested.

On benchmarks, the overall view was cautious. Final testing requires use-case consideration. General capability testing is still useful for domain-relevant capabilities such as number manipulation and tabular data understanding. But constructing tests fast is more important than benchmarking, especially because if testing is slow, it may lead to circumvention of MRM. Benchmarks were seen as useful for minimum functionality testing, but limited in showing reasons for failure, worst-case outcomes, or reliability. As one panellist noted, not all harms are equal, and not all metrics are equally informative.

Audience questions focused on 3rd party risks and accountability. If system trajectory sits with the vendor, firms may be limited to input-output testing. Questions were also raised about whether vendor evidence is relevant to a specific use-case and reinforcing the need for trace for auditability. On accountability, there was agreement that people remain accountable, but the harder question in multi-actor environments is which human or which vendor.

The closing message was pronounced: business pressure is moving faster than governance capacity to manage risks. The response cannot come from individual institutions acting alone—there is a need to work collaboratively, share insights, learn from lessons, and move forward.

Some questions for the panel that could not be accommodated due to time constraints included: differences in principles for pre-deployment testing and ongoing monitoring, and on the circular relationship between risk tiering and testing rigour.

Breakout Sessions

Workshop participants split into two groups based on their interests and engaged in round-table discussions.

Session 1: Evaluation Methods

The first breakout session focused on evaluation methods and was moderated by Turing's Christopher Burr.

On measuring utility, participants discussed what utility means for an AI agent in financial services, and for whom. Utility was linked to the performance indicators a firm seeks to improve, as well as to the fact that an agent is typically part of a broader workflow. It was considered alongside system stability and reliability. Participants also reflected on whether utility should be assessed at the level of the agent, the

overall system, or specific components, depending on the role of the agent. Measures discussed included cost reduction, savings in time and money, productivity gains, and revenue generation. Participants also highlighted the importance of considering short-term and long-term costs, including opportunity costs, and of comparing performance against a baseline, for example relative to human performance. Where benefits cannot be demonstrated clearly against the costs of inference, governance, and controls, systems may ultimately face decommissioning.

When the discussion turned to the evaluation of multi-agent systems, some participants questioned their relevance in practice, given the challenges of making such systems work, their utility relative to simpler approaches, and the technological trend toward increasing capabilities within single models.

On human versus AI evaluation, the status of human evaluation as a gold standard was challenged. Participants noted the extent of disagreement among human evaluators and the reliance on detailed instructions, explicit rules, and enumerated edge cases to obtain consistent human judgments. It was suggested that, if a similar approach is applied with instruction-following AI judges, that may offer greater auditability than human evaluation. At the same time, limited human evaluation was still seen as important for calibration and for identifying cases where disagreement itself is informative.

For black-box models, participants observed that there may be no practical alternative to brute-force input–output testing. In such cases, the overall approach to evaluation was framed as needing to remain risk-based.

On benchmarking, participants discussed Goodhart's law, with examples of bias testing and the use of demographic proxies. Benchmarks were described as needing to be non-stationary, reflecting changes in systems, environments, and user behaviour over time. Identified gaps included the difficulty of capturing variation in how users interact with LLM-based systems and the need for benchmarks that account for multi-turn interactions.

There was significant interest in testing for systemic risks and other evaluation-related issues, which could not be covered in detail in the session due to time constraints.

Session 2: Connecting Evaluation with Governance

This breakout session, moderated by Turing's Lukasz Szpruch, focused on the challenges of governing emerging AI systems and linking testing and validation to existing risk management frameworks.

Participants first highlighted that teams now face too many metrics, too many dashboards, and not enough clarity on what counts as an acceptable outcome. First-line teams often do not know what evidence and documentation are needed. AI risks cut across functions such as cyber and legal, so governance requires coordination. Many organisations still do not have a standardised approach and are still pathfinding through a few use cases. In some firms, governance is still being built through working groups and discussions.

Some participants noted that unlike governance and testing approaches for software, GenAI is multi-purpose, can create risks across the firm, and can also drive shadow AI. Use cases can shift after

deployment, and different parts of the business can have different interpretations of risk appetite. One approach discussed was to start in a more restrictive setting and relax it as confidence builds.

On risk appetite, participants discussed whether firms have clear statements that apply to AI. Some described appetite mainly in terms of business impact. Others said broad statements are not enough. Other approaches were discussed, including low appetite for some risks, zero tolerance for others, and explicit limits by model rating, tolerance period, etc.

On testing and validation, some participants said the issue is not a lack of metrics but deciding what to do with them. Questions included which metrics matter, when enough testing has been done, and who should decide across issues such as bias, security, and harassment. In systems combining AI and humans, testing becomes more complicated: risks may be delegated to humans, or metrics may exist but not be used effectively. One framing offered was that the validator's job is to try to break the model. Participants also noted that too many metrics bring monetary cost and environmental trade-offs, and that validators are often cautious because the risk comes back to them.

The session also covered incident sharing and third-party risk. Participants supported more systematic sharing of incident information and risk typologies so firms can learn from each other's mistakes. Third-party risks were described as one of the biggest risks and internal models were seen as relatively easier to manage. With external models and tools, firms may have limited information and limited leverage.

The discussion also briefly touched on issues such as regulatory policies and geopolitical competition.

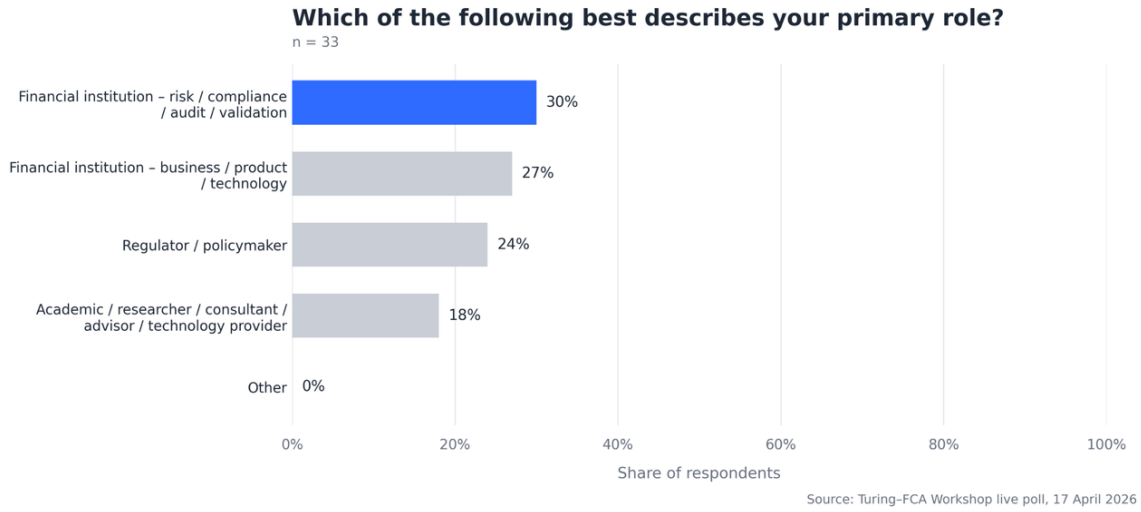
Reflections and Next Steps

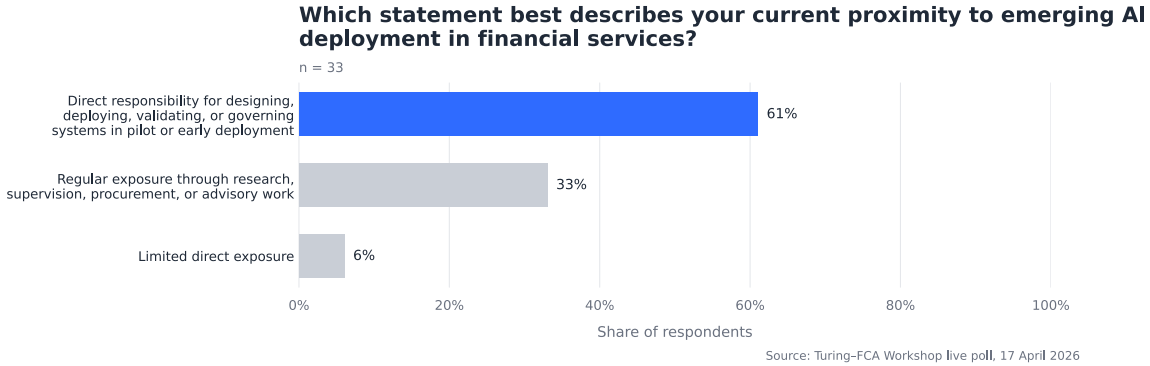
The discussion in this workshop benefited from the depth and diversity of expertise in the room. Participants drew on experience across deployment, validation, governance, and research, engaging thoughtfully with both technical detail and real-world constraints. Opportunities for this kind of cross-sector exchange remain limited, yet they are particularly important in a field evolving as rapidly as AI in financial services.

The organisers extend their sincere thanks to all speakers, panellists, moderators, and participants for their time, insights, and openness. Several important topics could not be explored in full within the time available, but this serves as a motivation for continued engagement.

Feedback on both the workshop and this readout is warmly encouraged, including suggestions for future themes and formats. We look forward to building on this initial workshop and fostering further opportunities for collaboration and shared learning in this space.

Appendix A


Which of the following best describes your primary role?
n = 33
Financial institution – risk / compliance / audit / validation
30%
Financial institution – business / product / technology
27%
Regulator / policymaker
24%
Academic / researcher / consultant / advisor / technology provider
18%
Other
0%
0%
20%
40%
60%
80%
100%
Share of respondents
Source: Turing–FCA Workshop live poll, 17 April 2026


Which statement best describes your current proximity to emerging AI deployment in financial services?
n = 33
Direct responsibility for designing, deploying, validating, or governing systems in pilot or early deployment
61%
Regular exposure through research, supervision, procurement, or advisory work
33%
Limited direct exposure
6%
0%
20%
40%
60%
80%
100%
Share of respondents
Source: Turing–FCA Workshop live poll, 17 April 2026