Building Trustworthy AI Frameworks in Financial Institutions

SFTI Roundtable at the SWISS AI SUMMIT 2025

Report

This report summarizes the essential insights and strategic discussions derived from the Roundtable on "Building Trustworthy AI Frameworks in Financial Institutions." The event was organized by SFTI at the SWISS AI SUMMIT in Zurich that took place on 17 November 2025.

The core objective was to create an open platform for exchange regarding strategic approaches to responsibly and effectively utilize Artificial Intelligence (AI) and developing a suitable governance framework for that purpose. The discussion featured leading experts from banks, insurers, academia, and leading AI companies and attendance was very high.

I. Governance Structure and the Necessity of Interdisciplinary Collaboration

The majority of the participants agreed that effective AI governance requires a holistic, risk-based approach integrated into the organization and built upon existing structures.

1. Strategic Integration and Holistic Governance

All experts emphasized that AI governance should be integrated into the organization's existing risk governance landscape, drawing on established risk management frameworks, particularly since many AI tools entail certain risks, especially regarding data that are already addressed by data governance and information security frameworks (e.g. Data Protection Impact Assessment).

However, as these existing risks may be amplified or even extended by AI, the participants cautioned against viewing AI merely as an IT tool manageable through standard processes – like any other type of IT outsourcing. Rather, they underscored that effective AI governance requires a broader, more adaptive perspective that takes into account these amplified or extended risks. To this end, relevant risk type owners must be identified and upskilled in order to be able to identify and assess any "new" risks in their area.

In order to make sure that the AI governance involves all relevant risk type owners and by doing so enables holistic assessments, a coordinating function within the organization was considered critical. As to the question of where the responsibility for this function should lie within an organization, compliance teams were broadly viewed as a suitable body. At the same time business involvement should not be underestimated and a robust AI governance framework has to be interdisciplinary, to ensure that all stakeholders are involved and all risks are addressed.

In the end, frameworks must be tailored to each target operating model and regularly updated to reflect technological progress and evolving regulation. At the same time, strategic foresight is necessary. This includes mapping the full lifecycle of each application, including post-implementation guidelines. This ensures that the assumptions behind deployment remain valid and reliable.

2. Interdisciplinary Collaboration: From Hurdle to Success Factor

Interdisciplinary collaboration, once identified as a major obstacle in the 2024 report, was recognized in 2025 as a critical success factor. Participants agreed that early involvement of all relevant functions – especially Business, Legal, Risk, and IT – is essential.

The participants noted that success rates for projects advancing beyond the Proof of Concept (PoC) stage improved significantly due to the earlier inclusion of key stakeholders. Without such early collaboration, pilot projects tend to fail.

3. AI Risk Categorization

Given the growing integration of AI functionalities across various services, AI tools or use cases should be categorized based on key risk factors such as the degree of automation, the likelihood of malfunction, potential business or customer impact, and legal or regulatory exposure. Medium- and high-risk tools or use cases require comprehensive assessment and more extensive testing before deployment.

Participants highlighted that risk assessments should not rely solely on external regulations like the EU AI Act but must consider each institution's specific risk profile and operational context as well as sector-specific regulation. Several experts proposed a pragmatic, staged approach: organizations could first focus on low-risk, high-value applications – such as search, information extraction or process automation – to foster internal experience and refine the designed governance practices before tackling more complex or sensitive use cases. They also emphasized that building a trustworthy AI governance framework is an iterative process and none of the participants have a blueprint ready that could be used or scaled for other companies in the financial sector just yet.

II. Implementing Abstract Principles and Scaling AI

One of the central challenges when building trustworthy AI governance frameworks lies in translating abstract governance principles – such as transparency or explainability – into practical, operational measures. These principles are rooted in frameworks like the AI Act, the OECD AI Principles, and the FINMA Guidance 08/2024.

1. Practical Transparency and Explainability

Participants highlighted that transparency must be meaningful and accessible to humans, not merely a technical logging exercise. There was broad agreement that no perfect solution for audit logs currently exists, and that true transparency – and what it entails – depends on continuous dialogue among diverse stakeholders within an organization. Since model evaluation rarely yields a single correct answer and no one-size-fits-all solution has emerged, explainability must remain a collaborative and evolving effort. In this context, open source is likely to play an increasingly important role. To enhance effective coordination and mitigate risks of duplication or conflicting AI actions, the participants emphasized the critical importance of maintaining a transparent, comprehensive inventory of all AI tools and use cases deployed across the organization. Such an inventory supports clear accountability, facilitates governance oversight, and enables ongoing compliance monitoring. Furthermore, it enables knowledge sharing and broader reuse of the existing capabilities. Standardization such as the ISO 42001 standard will also play an important role in implementing the principles defined in existing regulations and one of the participants shared some insights as to what such a certification under the ISO 42001 standard entails.

2. Transition from PoC and Scaling

Moving AI initiatives from Proof of Concept (PoC) to full production represents a critical and challenging phase. This transition requires clear objectives, deep understanding of the prerequisites, well-defined success metrics, and early alignment among all stakeholders to ensure the solution aligns with business goals and technological requirements. Larger organizations tend to run fewer pilots but achieve higher success rates due to more structured processes and governance frameworks, whereas smaller firms often conduct numerous pilots but face difficulties scaling effectively. A pragmatic approach for smaller companies is to adopt proportionate governance models supported by robust testing suites and benchmarking tools.

III. Agentic AI under Control

Discussions reaffirmed the importance of maintaining strict control over autonomous (agentic) AI systems. Oversight must remain central, particularly for high-stakes decisions.

1. The Enduring Role of Human Oversight

The contributors agreed that full automation of high-level decisions is unacceptable due to the potential risks that would entail, and the duty of care and regulatory requirements financial institutions have to comply with. Human involvement must always be part of the decision loop to align with both institutional principles and regulatory expectations. The participants also emphasized that regulators will continue holding individuals – not AI systems – accountable for decisions made using flawed data, underlining the need for precise knowledge of data sources and quality.

2. Control Mechanisms

One of the participants shared an incident where an AI-based model became stuck in a loop and generated significant costs, which ultimately provided valuable insights into how this rapidly evolving technology behaves in real-world conditions. The key takeaway was that agentic AI can scale output dramatically but also multiplies potential errors, demanding a corresponding increase in control. Such experiences, in turn, help institutions refine their monitoring and oversight as they bring new applications into production, turning early challenges into opportunities for stronger, more resilient systems.

To mitigate the risks posed by autonomous agents, organizations should employ safeguards such as controlled sandboxes, real-time monitoring, and kill switches capable of detecting and responding to deviations or “drift” in system behavior. Many participants referred to AI (agents) as new employees, they have to be instructed (or prompted) correctly, given the right context and supervised at every step so that they can learn from the experienced managers and execute the tasks in a way that matches the real-world expectations.

Conclusion

The overarching message remains clear: None of the participants expects AI to replace human judgment completely. Its successful and foremost trustworthy integration relies on robust, interdisciplinary governance that guarantees a holistic risk assessment by upskilled subject matter experts as well as adequate safeguards such as human oversight in high-stakes decision-making processes. Future progress in the financial sector will depend on maintaining this equilibrium of balancing innovation with clear governance and ongoing investment in AI literacy.

Beyond safeguarding compliance, this approach opens a strategic horizon: organizations that embed trustworthy AI and strong governance not only mitigate risks but also unlock innovation, strengthen stakeholder confidence, and position themselves for a sustainable competitive advantage in an increasingly AI-driven market.

Moderation*:* Rehana Harasgama

Report*:* Jonas Tresch