Mila x Finance : The Era of Agents, Risk, and Consumer Protection

Summarized findings from the financial industry

Mila

About this report

On March 31, 2026, Mila — Quebec Artificial Intelligence Institute hosted a collaborative World Café-style event. The gathering brought together approximately 30 industry professionals from around 15 Canadian financial institutions. The primary goal was to foster dialogue between Mila and the financial sector, addressing shared challenges and aligning Mila's research objectives with market realities. The focus was on developing robust financial AI, particularly in the areas of risk management and consumer protection.

Discussions were structured around five thematic roundtables:

I. AI GOVERNANCE: FRAMEWORK AND REGULATORY LANDSCAPE

II. AI RISK ASSESSMENT AND OVERSIGHT

III. TECHNICAL SAFETY, GUARDRAILS, AND MEASUREMENT

IV. AI IN RISK MANAGEMENT: FRAUD DETECTION USE CASE

V. THE AGENTIC FUTURE

Each table was facilitated by a member of Mila's applied research team and an academic researcher from the institute. This report summarizes the critical points and industry challenges covered during these discussions.

Acknowledgements

This report represents a collaborative effort by Mila – Quebec Artificial Intelligence Institute, and leading representatives from the financial industry of Quebec. We thank the following members of Mila for their essential presence, expert guidance, facilitation of the discussions, and valuable contributions to the drafting and development of this report: **Istabrak Abbes**, Master's Research Candidate, Université de Montréal, **Arsène Fansi Tchango**, Manager, Applied Machine Learning Research, **Simona Gandrabur**, Head of AI Safety Studio, **Prakhar Ganesh**, PhD Candidate, McGill University, **Gaétan Marceau Caron**, Senior Director, Applied Machine Learning Research, **Philippe Martin**, PhD Candidate, Université de Montréal, **Maryam Molamohammadi**, Senior Scientist, Responsible AI Research, **Mahta Ramezanian**, PhD Candidate, Université de Montréal, **Shalaleh Rismani**, Postdoctoral Researcher, McGill University, **Adam Salvail**, Manager, Applied Machine Learning Research, **Elnathan Tiokou**, Master's Research Candidate, Polytechnique Montréal, and founder of the Mila startup Vraust.ai.

The event was organized, and this report was written, under the guidance of **Rheia Khalaf**, Director of Partnerships at Mila.

We wish to express our deep gratitude to the industry participants who shared their candid perspectives and valuable hard won practical insights. Participants included representatives from banks, insurers, asset managers, and regulatory agencies, all committed to balancing the risk and opportunity inherent in the continued advancement of AI within financial services. Participating organizations included: Autorité des marchés financiers (AMF), National Bank of Canada, BNP Paribas, Desjardins Group, Finance Montréal, iA Financial Group, La Caisse de dépôt et placement du Québec (La Caisse), Manulife, Office of the Superintendent of Financial Institutions (OSFI), RBC Borealis, Société Générale, and TD Bank Group.

Note: The content of this report was refined and improved using generative AI models, with careful human oversight and final validation.

Foreword

The adoption of AI systems is evolving from experimental pilots to a fundamental restructuring of the corporate environment. These systems now operate as semi-autonomous coworkers, not just software tools. However, many organizations that are rapidly adopting these technologies face the strategic challenge of creating mechanisms and systems around these technologies that deliver cohesive business value.

This rapid integration of these systems creates a critical tension with the need for human oversight. Without robust management, companies are exposed to significant operational and reputational risks as AI takes on high-stakes decision-making. To counter this, organizations are embedding security and governance directly into their workflows, utilizing dynamic guardrails that automatically trigger human intervention when specific risk thresholds are exceeded.

The financial sector, like many other highly regulated industries, is grappling with common challenges in the rapid adoption of AI. To address these issues, Mila recently convened a strategic discussion bringing together experts from the financial industry and regulatory bodies. The current pace of technological change and AI adoption necessitates renewed collaboration between the public, private, and academic sectors. The goal is to transcend the pilot trap and embrace a more AGILE framework (Awareness, Guardrails, Innovation, Learning, and Ecosystem Resiliency), as introduced in the Financial Industry Forum on Artificial Intelligence (FIFAI) II report[1]. The discussion's objective was to move beyond theoretical issues and tackle the practical difficulties of integrating AI into regulated environments:

→ **Aligning Innovation with Market Demands:** Focus on discussing solutions that strengthen the resilience of the financial system and enhance robust consumer protection, while directly addressing high-impact constraints such as auditability and data sovereignty.

→ **Fostering Interdisciplinary Collaboration:** Transform governance from a late-stage bottleneck into an accelerator for secure innovation, ensuring a foundation for scalable, trustworthy, and compliant AI deployment.

This report outlines the specific AI risks and opportunities identified during the discussions. It is structured to reflect the necessary sequence for successful AI operationalization, moving from high-level policy to implementation and future vision. It begins by examining the core operational rules and regulatory context required for **AI Governance**. Next, it addresses the structural foundations for **AI Risk Assessment and Oversight**. This structural foundation sets the stage for the **Technical Safety and Guardrails** section, which explores the real-time enforcement mechanisms required for policy compliance. These combined principles are then tested in **AI for Risk Management**, detailing the core financial use case of fraud detection and its specific constraints. Finally, the report concludes by looking ahead at **The Agentic Future**, analyzing the complexities and the controlled, incremental path required for adopting autonomous AI at scale.

[1] FINANCIAL INDUSTRY FORUM ON ARTIFICIAL INTELLIGENCE II, March 2026 report
https://globalriskinstitute.org/publication/fifai-ii-ai-risks-and-opportunities/

Executive Summary

The scaling of AI in the financial sector is fundamentally constrained not by technology, but by a lack of operational maturity in governance, data readiness, and defined accountability. Organizations must prioritize building robust foundational infrastructure, including risk-based governance and technical safety systems, to avoid accumulating governance debt and move past the pilot phase.

Key Findings Across AI Operationalization and Risk Management:

I. AI GOVERNANCE: Financial AI governance is a mandatory, risk-based mandate requiring embedded, foundational data governance to navigate complex global regulations and overcome the scaling gap from Proof of Concept (PoC) to production.

II. AI RISK ASSESSMENT: Oversight must continuously monitor the four pillars of risk (Reliability, Ethics, Data/Privacy, Security/Safety), manage inherent trade-offs (fairness paradox), and address risks from third parties and organizational maturity gaps.

III. TECHNICAL SAFETY: Technical guardrails are essential, non-optional safety architecture in high-stakes environments, demanding domain-specific benchmarks and balancing safety improvements with latency costs, especially in complex agentic systems.

IV. AI IN RISK MANAGEMENT: Fraud detection is the top AI use case in banking, challenged by data scarcity, high false-positive rates (hallucination trade-off), and sometimes finding simple rules more effective than complex AI models.

V. THE AGENTIC FUTURE: AI agent deployment is currently limited to internal productivity PoCs, blocked from scaling by reliability concerns, lack of traceable governance, and the imperative to maintain human judgment as the final authority.

Operationalization and Risk Management

We are in a rapidly evolving environment, as new models, new risks, and threats appear every day. The need for a report detailing current shared practices is critical to ensure that organizations can proactively adapt to the speed of change, maintaining both security and competitive advantage. The below sections describe shared practices as of the date of writing of this report, serving as a vital benchmark in this dynamic landscape.

I. AI GOVERNANCE: FRAMEWORK AND REGULATORY LANDSCAPE

The financial industry is currently engaged in a high-stakes race to integrate artificial intelligence, but the core challenge lies not in the technology's capability, but in establishing the necessary rules of engagement. This journey of operationalizing AI governance is unfolding across a landscape defined by rapid experimentation, complex regulatory pressures, and intense internal demands to prove value.

1. Contextual and Risk-Based Frameworks

→ **AI Governance is Regulatory:** Financial institutions, as a highly regulated sector, face unique pressures in establishing AI governance. Unlike less regulated industries, they cannot treat AI governance as an optional layer; it is a regulatory mandate interwoven with existing compliance regimes (e.g., anti-money laundering, consumer protection, data privacy). This environment demands a more conservative, risk-averse approach, prioritizing auditability, clear accountability, and human-in-the-loop oversight, especially for high-impact use cases. The challenge is not just compliance, but translating fragmented global regulations (like the EU AI Act and local data laws) into a unified, actionable internal framework that still allows for necessary innovation without accumulating insurmountable technical or governance debt.

→ **Data and AI Governance:** A recurring industry insight is that organizations frequently lack full readiness for AI governance because they are still addressing foundational challenges in data governance. Issues of data quality directly impact downstream AI systems, underscoring that AI governance is inextricably linked to and fundamentally dependent on robust data governance, as poor data practices undermine accountability. This is intrinsically linked to the need for improved documentation and recordkeeping, which are essential for traceability, regulatory compliance, and understanding failures over time.

→ **Shift to Contextual, Risk-Based Frameworks:** There is a strong move across institutions away from rigid, one-size-fits-all governance toward contextual, risk-based frameworks. Oversight is increasingly tailored to the specific AI use case, meaning that the distinction between high vs. low risk determines the level of scrutiny and speed of deployment. For instance, the risks associated with AI vary significantly based on factors like whether the system is client-facing or internal, the scale of its potential impact, and whether it is self-owned or relies on third-party vendors. This context-dependent approach directly informs AI risk management and mitigation strategies. Since the nature of the AI use case dictates the type and severity of the risk, mitigation efforts must be highly specific. For example, the risk of data leakage in a client-facing generative AI product requires different controls than the risk of bias in an internal loan approval model.

→ **Governance as an Accelerator:** The key to success is to embed AI governance into existing processes rather than layering entirely new frameworks. When governance is introduced late, it is often perceived as the police and becomes a blocker to innovation. Conversely, when governance is integrated early into development, risk, and compliance workflows, it can accelerate adoption. Institutions are increasingly recognizing that governance must be practical, embedded, and aligned with existing systems to be effective.

2. Governance Challenges in Scaling from Proof of Concept (PoC) to Production

→ **Bridging the Gap: From PoC to Production:** A major structural gap exists between experimentation and scaling. A large share of initiatives remain in PoC, and the transition from PoC to production is often slow and resource-intensive. Institutions are trying to manage this by relaxing constraints during experimentation (e.g., sandbox environments) and reintroducing strict controls when necessary. Guardrails are also being designed horizontally across use cases, rather than being re-evaluated from scratch each time. However, even successful PoCs frequently stall when faced with integration and scaling requirements.

→ **The Budgeting Dilemma:** Funding AI projects, especially PoCs, remains structurally difficult. The lack of clear ROI for experimentation raises questions about who funds it and which business unit is responsible for its long-term governance and accountability. As a result, promising ideas may stall before reaching production.

→ **The Unresolved Metric of Success:** There is no clear consensus on standard success metrics for AI governance. Institutions are using a mix of approaches, including efficiency gains (time and cost savings), error rates (human vs. AI comparison), and risk indicators such as data breaches or reputational impact. Post-deployment, there is strong emphasis on monitoring plans, drift detection, and continuous threshold adjustment, but no unified framework.

3. Global Regulatory Fragmentation and Translation

→ **Global Regulatory Fragmentation and Translation:** The regulatory landscape is fragmented and rapidly evolving, complicating matters for organizations navigating multiple overlapping regimes such as the EU AI Act, General Data Protection Regulation (GDPR), Law 25, and emerging Canadian and U.S. guidelines. This necessitates hybrid internal frameworks, often driven by the most stringent jurisdiction, and requires institutions to constantly monitor updates and translate complex legal requirements into business and technical implementation. The environ-

ment remains non-standardized: Europe favours prescriptive approaches, North America offers flexibility, and Asia is still experimenting. While regulatory bodies demand explainability, accountability, and traceability, they often do not prescribe specific implementation methods. Consequently, financial institutions operate in a dynamic setting where core expectations are established, but the precise path to compliant and scalable execution is still being defined by evolving industry practices.

II. AI RISK ASSESSMENT AND OVERSIGHT

To effectively scale AI beyond initial pilot phases, robust auditing and governance are essential. This necessitates moving past simple compliance checklists and adopting contextual, risk-based frameworks for AI risk assessment.

1. AI Risk Assessment Frameworks

→ **Primary Pillars of Risk Assessment:** These frameworks shift the focus from conventional IT security concerns to the wider socio-technical impacts of AI, commonly categorizing risks into four primary pillars.

- **Technical Reliability:** Accuracy and robustness are paramount. This includes preventing hallucinations and ensuring the AI performs reliably even when exposed to real-world data that deviates from its training set. Given that generative AI relies on vast, unstructured datasets (scraped text, images, and code), the risk of «garbage in, garbage out» is significant. Data lineage is the framework used to ensure the model's «knowledge» is traceable, verifiable, and protected against malicious injections or significant decay.
- **Ethics and Fairness:** This area focuses on identifying algorithmic bias that could lead to discrimination. It also prioritizes explainability, ensuring that AI decisions are not black boxes but can be understood by humans.
- **Data and Privacy:** This addresses the fuel of AI. Key risks include data leakage (where the model may reveal sensitive information), lack of user consent, and the use of copyrighted material, which can create legal liability.
- **Security and Safety:** This goes beyond traditional cybersecurity threats to include adversarial attacks, such as prompt injection, as well as physical safety risks, particularly in contexts like automated manufacturing systems or autonomous vehicles.

→ **Complexity in AI Risk Management and Mitigation:** AI risk management is inherently complex because it requires balancing deeply interconnected and often contradictory objectives, where optimizing for one safeguard can inadvertently weaken another. This is most visible in the fairness paradox, where strict adherence to data privacy principles, such as removing personally identifiable information (PII) to protect identities, directly undermines ethical fairness by stripping away the demographic data necessary to audit the system for bias. Beyond these internal trade-offs, the landscape is further complicated by the 'black box' nature of many models, where transparency is often sacrificed for higher predictive accuracy. Because the risk areas of privacy, fairness, security, and performance do not exist in isolation, mitigation is not a simple checklist but a continuous act of calibration; a fix for one vulnerability can create a ripple effect that introduces legal, social, or technical liabilities elsewhere.

→ **Continuous Risk Management Cycle:** A robust risk management framework necessitates a continuous cycle of monitoring, measurement, and adaptation. This ensures controls remain proportional to the dynamic and context-specific risks posed by the AI system throughout its entire lifecycle. Generative AI is prone to drift, where the model's outputs degrade or shift over time as it interacts with new data. This reinforces the need for organizations to have continuous monitoring systems to validate that the AI is still operating within its original safety and accuracy parameters.

2. Data Sovereignty and Third-Party Risk

→ **Data Sovereignty Concerns:** Relying on external AI tools raises significant concerns regarding data sovereignty. When third-party providers process data, institutions face risks related to where the data is stored, which legal jurisdiction applies, and the potential for unauthorized access or use of sensitive information by the external provider. This loss of control over data location and access is a key driver for developing internal AI solutions, despite performance trade-offs.

→ **Third-Party Trade-Offs:** A major concern is the trade-off between third-party AI tools and data sovereignty. External tools are often more powerful and user-friendly, but introduce risks related to data leakage, loss of control, and lack of visibility. As a result, institutions are pushing to develop internal tools, even if they are less performant. In practice, this creates shadow behaviors, where employees use personal accounts or unofficial channels to access better tools.

→ **Vendor Model Risk Accountability:** A specific, escalating challenge is the Model Risk Management (MRM) of third-party embedded generative AI tools. While financial institutions remain fully accountable for the MRM of these models, many vendors refuse to share sufficient details of their risk assessment or internal workings. This lack of transparency necessitates that AI governance policies establish clear frameworks and contractual requirements to mandate information access and risk data-sharing from vendors.

→ **Regulation and Non-Technical Solutions:** Effective AI governance acknowledges the limits of purely technical solutions. Instead, governance must incorporate non-technical measures, such as stronger contracts, better documentation (e.g., model cards[2]), and clearer expectations from third-party providers. Since smaller companies often lack leverage over large, monopolistic AI providers, regulations play a critical role in setting industry standards. Uncertainty persists around existing regulatory frameworks and the evolving division of roles between developers and deployers. Overall, effective AI governance requires coordination between internal practices and external regulatory structures.

3. Organizational Structure and the Lines of Defence

→ **The Three Lines of Defence Model (3LOD):** Implementing effective AI governance often involves three lines of defence, typically including developers, an internal audit team, and a separate independent audit team. A common institutional challenge is the lack of sufficient AI expertise across these three lines of defence, resulting in gaps in oversight. In some instances, audit teams involve developers in auditing their own systems, which raises concerns about independence and potential bias or contamination.

[2] https://huggingface.co/docs/hub/model-cards

→ **Problem of Many Hands:** This complexity contributes to the problem of many hands, where diffused responsibilities across multiple teams make accountability unclear. Effective governance necessitates interdisciplinary collaboration and a cultural shift where failures are accepted as opportunities to improve structures, rather than blaming individuals. Proactive risk management, rather than reactive "emergency room" responses, requires engagement from senior leadership.

→ **The Efficiency-Oversight Tension:** The core value proposition of AI efficiency is continuously tempered by the non-negotiable demand for human oversight. In high-risk domains like anti-money laundering, human judgment is essential for accountability, meeting regulatory requirements, and mitigating risk. Effective risk mitigation requires a proactive approach, integrating thorough legal analysis and human oversight throughout the AI development lifecycle. However, the efficacy of human oversight remains questionable, and this requirement introduces operational friction. This friction can diminish AI's expected efficiency gains and lead to operational, administrative, and financial delays. Conversely, rapidly deploying new AI features without adequate legal review often necessitates costly removal or rework down the line.

→ **Navigating Internal Dynamics: Literacy, Alignment, and Change Management**: Internally, organizations face significant change management pressure driven by misalignment. While executives are eager to capture return on investment and driven by the urgency to adopt AI ahead of competitors, employees struggle with both curiosity and concern over job transformation. Middle management is tasked with operationalizing AI and balancing these competing expectations, highlighting a critical need for enhanced AI literacy, training, and clear communication, as well as recognition that adaptability is becoming a core skill.

III. TECHNICAL SAFETY, GUARDRAILS, AND MEASUREMENT

The clarity of regulatory expectations (explainability, accountability, and traceability) contrasts sharply with the non-prescriptive implementation methods. This dynamic shifts the focus from high-level governance to the critical, real-time technical mechanisms required to enforce organizational and regulatory requirements. This mechanism is the guardrails, an independent safety system that filters AI inputs and outputs in real time against specific policies to prevent policy violations.

1. Designing and Enforcing Real-Time Guardrails

→ **The Evolving Role of Guardrails in Finance:** In high-stakes sectors such as finance, guardrails can no longer be treated as optional add-ons to language models. They must be designed as part of the system architecture, monitored over time, and evaluated against realistic failure modes. Furthermore, the same model may require different safety configurations depending on the use case, level of autonomy, and user population. This shifts the role of guardrails beyond simple refusal systems [3]. In practice, guardrails must help enforce organizational boundaries, detect policy violations, and support compliance requirements while remaining compatible with enterprise workflows. Administrative controls and technical controls should be viewed together rather than separately.

[3] A simple refusal system in machine learning is designed to identify and decline harmful, unsafe, or out-of-scope queries. These systems are crucial for safety, ensuring that models, such as chatbots, do not generate dangerous, illegal, or unethical content.

→ **Managing Agentic Systems' Complexity:** With agentic systems, the concern is no longer limited to whether one answer is correct. The system may browse, trigger tools, write data, scrape content, or operate over internal infrastructure, which creates a larger attack surface. Prompt injection, web injection, excessive permissions, and poorly scoped human-validation loops present concrete concerns. Complexity itself becomes a risk multiplier: as systems become more autonomous, failure analysis, responsibility assignment, and change management become harder.

→ **Integrating Industry-Specific Guardrails:** The challenge is not only harmful content generation, but also unauthorized access, propagation, or transformation of sensitive information. There is a strong need for domain-adapted guardrail benchmarks for finance, specifically addressing privileged, sensitive, or regulated information leakage, privacy and Personally Identifiable Information (PII) handling, as well as unsafe investment advice. These benchmarks are used to evaluate the effectiveness, reliability, and speed of the guardrails. The path forward involves modular guardrail architectures that combine fast specialized classifiers, configurable policy libraries, and the selective use of LLM judges or human review. Continuous evaluation and multilingual robustness are major areas for improvement, alongside designing safety-aware workflows that estimate risk before and after an action, not only after an incident occurs.

2. Evaluation Challenges and Latency Trade-Offs

→ **Tailoring Benchmarks to Domain Needs:** Evaluation is one of the most urgent needs. A guardrail is only meaningful if its performance can be measured in a way that matches the deployment context. A useful shift goes towards tailored benchmarks, dynamic real-time evaluation, monitoring dashboards, and use-case-specific metrics rather than generic leaderboard-style comparisons. The financial industry lacks widely shared public benchmarks because realistic datasets are difficult and costly to construct, especially when privacy and confidentiality constraints apply. As a result, custom evaluation pipelines, often requested by clients or built internally, appear to be the practical path forward.

→ **Tension Between Safety and Latency:** large language model-as-a-judge approaches, an AI evaluation technique where a high-performing Large Language Model is prompted to assess the outputs of another model or system, are seen as useful for verification and post-hoc assessment, but their cost and latency overhead, in some cases, roughly double or triple response time. That makes them attractive for high-risk decisions, audits, or asynchronous review, but less suitable as a universal runtime solution. This leads to a broader architectural question: which risks should be handled by lightweight specialized detectors, which by LLM-based adjudication, and which by human escalation? It can be argued that correctness often matters more than speed in critical scenarios, but only up to a point; institutions will still need tiered safety stacks that allocate computational budgets according to risk severity.

3. Measurement and Placement Best Practices

→ **Placement Strategy:** Some harms are best addressed at the input stage, for example when a topic should not be engaged at all, while others are better addressed at the output stage, where the system must decide how to respond safely within an allowed domain. For enterprise assistants or customer-facing systems, risk can accumulate across turns rather than appear in a single utterance. In this case, static single-turn filtering is insufficient; streaming and multi-turn settings (maintaining conversational context over multiple interactions) are of better use. In general, placement may vary by domain: for some sensitive applications, post-generation checking may be more effective than pre-generation prompt filtering alone.

→ **Measurement Focus:** Key practices for guardrail measurement include focusing on refusal-to-respond rates, developing domain-specific attack taxonomies, conducting adversarial prompt testing, and monitoring behavioural changes in AI systems over time.

→ **Robust Protocols:** Since AI is non-deterministic, measurement frameworks must recognize that a different answer is not always a bug. Evaluation efforts must therefore focus on distinguishing harmless variation from meaningful safety drift. This necessitates adopting robust, longitudinal evaluation protocols instead of relying on one-time benchmark runs.

IV. AI IN RISK MANAGEMENT: FRAUD DETECTION USE CASE

A practical integration of AI sits within the financial sector's core function: risk management. This domain encompasses integrating AI into existing processes to manage risks such as pre-identified threats, like fraud.

1. Core Use Case: Fraud Detection and Operational Constraints

In the context of AI for risk management, fraud detection represents the most prominent use case in banking. Fraud department budgets have drastically increased over the past decades in many worldwide banks. With the pace of AI development, fraudulent activities have become significantly more efficient, including the falsification of identity documents, trustworthy phishing emails, and cybersecurity breaches. Financial institutions must carefully incorporate AI tools for fraud mitigation while staying aware of the risks that arise with them.

→ **Data Scarcity and Anomaly Detection Techniques:** Fraud detection is often hindered by a weak supervision problem, where labeled fraudulent data (true positive fraud labels) is rare compared to legitimate transactions, and false negatives are dominant. Several solutions for fraud detection are being utilized:

- **Synthetic Data:** Artificially creating fraud scenarios to train models when real-world examples are insufficient. The synthetic generation of financial behavior data with properly defined scenarios benefits detection algorithms, as this newly generated data can be used in a supervised manner in score/classification algorithms that currently lack proper labeled data.
- **Anomaly Detection:** Utilizing unsupervised techniques like Isolation Forest[4], which are popular methods for the detection and scoring of abnormal behaviors. This helps identify outliers, such as unusual transaction patterns in accounts belonging to vulnerable populations (e.g., transactions by seniors over 80 at nightclubs), or suspicious activity related to money transfers and payments to at-risk countries and organizations.
- **Operational Advantages of New Agents:** New AI agents can offer advantages by identifying known fraud scenarios (typically derived from web content included in the LLM's training data) and automatically testing them, as well as applying different algorithms to generate client risk scores.

[4] Isolation Forest is an unsupervised machine learning algorithm specifically designed for anomaly detection. Unlike traditional methods that build a profile of 'normal' data to find deviations, it explicitly focuses on isolating individual observations.

→ **The False Positive Trade-Off and Simple Interventions:** In fraud detection, increasing a model's sensitivity to detect new fraud patterns often leads to a high volume of false positives (alerts incorrectly flagging legitimate transactions). This issue is distinct from a model's hallucinations (generating factually incorrect or nonsensical output), as false positives refer to misclassification errors within a defined scope. A high false-positive rate, such as a jump from 10% to 35%, is economically unsustainable, as it overwhelms human analysts with noise. To stay ahead of sophisticated fraudsters who use AI to eliminate traditional phishing signals (e.g. spelling errors), simple logic is often found to be more effective than complex AI alone. For instance, requiring an exact name-match for IBAN transfers can eliminate a vast category of payment fraud.

V. THE AGENTIC FUTURE

While AI agents hold significant promise for transforming financial operations, the adoption of these autonomous systems remains at an early stage. Most organizations are running proofs of concept (PoCs) rather than deploying agents in production, with current efforts tightly focused on internal tools and controlled actions.

1. The Agentic Future: Deployment and Scaling

→ **Use Cases Center on Productivity Gains:** Automating repetitive junior-level tasks can save an estimated 30–40% of time. Client-facing applications are limited to chatbots and guided workflows restricted to non-sensitive data. Agents are also used for decision support through qualitative analysis and structured reporting, as well as for workflow automation such as sector monitoring and standardised report generation.

→ **Risk Management for Agentic Commerce:** Eventually, purchases and financial transactions will occur end-to-end between authorized AI agents without human involvement. This autonomous environment creates an acute need for new risk management protocols, traceability, and defined liability, especially concerning transaction validation, fraud, and ensuring consumer protection when financial decisions are executed exclusively by agents.

→ **Several Barriers Prevent Broader Deployment:** Firms lack controlled environments with sufficient governance, monitoring, and rollback capabilities. Reliability concerns persist, including hallucinations and unstable outputs in high-stakes contexts. Internal expertise is limited, making it difficult to keep up with rapid vendor innovation. Organizations face a tradeoff between building in-house, which offers control but is slow, and partnering with vendors, which accelerates delivery but introduces auditability and dependency risks, as we have seen earlier. Regulatory constraints require validation, traceability, and clear accountability, while rapid vendor turnover increases operational complexity. Existing data architectures are often not ready for agent-based systems.

2. Governance and Technical Best Practices for Agentic Systems

→ **Continuous Monitoring of Inputs, Outputs, and Model Lineage:** Liability is currently assigned to users, though future regulation may shift responsibility toward model providers. Misunderstanding open-source licenses and model provenance also introduces legal risk.

→ **Technical Best Practices:** These include limiting agent actions to read-only or tightly constrained operations, implementing detailed observability with immutable logs[5], and deploying automated evaluation tools to detect hallucinations and ensure compliance. Data should be segregated, using synthetic or de-identified datasets in testing environments and strict access controls in production.

→ **Key Performance and Safety Metrics:** These include hallucination rate, action error rate, time to detect and fix issues, false positive and negative rates for risk detection, and the human effort required to verify outputs.

3. Major Risks and Path Forward

→ **Controlled Scaling:** The main risks involve reputational damage from client-facing errors, regulatory escalation due to insufficient auditability, and technical debt from fragmented vendor integrations. Progress requires a controlled, incremental approach: start with sandboxed environments, focus on narrow workflows with clear metrics, implement automated evaluation, and enforce vendor audit controls. This allows organizations to manage risk while preserving the option to scale successful pilots.

→ **The Human Frontier:** While AI agents will increasingly serve as productivity enhancers for financial experts, human judgment remains the final arbiter for high-stakes decisions. Employees must therefore undergo important training to measure the potential, limits, and associated risks of new AI agents. The future of finance lies in the balance between cutting-edge algorithmic detection and the preservation of human-centric legal and ethical standards.

[5] https://trainingcamp.com/glossary/immutable-logs/

Conclusion

The *Mila x Finance: The Era of Agents, Risk, and Consumer Protection* event underscored a critical consensus within the Canadian financial sector: the successful scaling of AI hinges on the maturity of governance, risk, and safety frameworks, not just on technological capability. The transition from isolated Proofs of Concept (PoCs) to enterprise-wide production requires a systemic shift, moving from siloed development to integrated, risk-based operationalization.

The discussions highlighted that effective AI adoption demands:

→ **Embedding Governance Early:** AI governance must be treated as a non-optional regulatory mandate, integrated early into the development lifecycle, and built upon a strong foundation of data governance to ensure accountability and auditability.

→ **Continuous and Contextual Risk Management:** Oversight frameworks must continuously monitor for reliability, ethical bias, privacy, and security risks, recognizing the inherent trade-offs (like the fairness paradox) as well as the complexity introduced by third-party vendors and the problem of many hands.

→ **Non-Negotiable Technical Safety:** Technical guardrails are essential architectural components for real-time compliance enforcement in high-stakes environments. They require domain-specific benchmarks and sophisticated strategies to manage the risk multiplier effect of agentic complexity without introducing excessive latency.

→ **Controlled, Incremental Scaling:** While the agentic future promises significant efficiency gains, its current deployment is limited by reliability and auditability concerns. The path forward requires a controlled approach, starting in sandboxed environments, focusing on narrow workflows, and ensuring that human judgment remains the final decision authority in high-stakes contexts.

Ultimately, the future of AI in finance lies in achieving a robust balance. It requires transcending the pilot trap by adopting an AGILE framework that prioritizes awareness, protective guardrails, and ecosystem resiliency. This collaborative effort between academia, Mila, industry, and regulatory bodies is vital to ensure that as AI evolves into a coworker, it remains trustworthy, compliant, and fundamentally protective of the consumer.

"*The Mila x Finance: The Era of Agents, Risk, and Consumer Protection* event shed light on the rapid evolution of AI and the new risks it brings to the financial sector. While fraud detection tools are advancing, fraudsters are now using AI themselves to bypass banking safeguards — a challenge that was virtually non-existent just two years ago. Through the roundtable discussions, we were able to exchange ideas with fellow experts on real-world internal applications and share what is actually working in the field of detection. With technology evolving so quickly, annual gatherings like this one are essential. They allow us to stay ahead of the curve and anticipate risks we had not yet foreseen."

- Philippe Martin, PhD Candidate, Université de Montréal, facilitator for AI in Risk Management: Fraud Detection Use Case roundtable discussion

Partner with Mila

Whether you are looking to launch a research collaboration, explore AI applications for finance and risk management, accelerate responsible AI adoption, or connect with leading AI talent, we want to hear from you.

By partnering with Mila, your organization gains access to:

World-Class Expertise: Collaborate with leading researchers in machine learning, generative AI, optimization, forecasting, and trustworthy AI.

Applied Innovation: Explore high-impact use cases across financial services, including risk management, fraud detection, climate finance, portfolio optimization, and compliance.

A Unique Ecosystem: Leverage Mila's connections with industry leaders, startups, regulators, and the broader AI community.

Talent & Capability Building: Engage with top AI talent through collaborative projects, workshops, and innovation programs.

Contact Mila's Partnerships team to explore how your organization can help shape the future of AI in finance.