Federated Multi-Agent Learning for Privacy-Preserving Financial Risk Management

Irmtraut Bolzmann
Department of Computer Science, University of North Texas, Denton, TX, USA.
irmtraut.work@unt.edu

Espartaco Rosas
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
espartaco894@ucf.edu

Karan Zachariah
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
karan29@oregonstate.edu

Yadavi Bhakta
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
helloyadavi@uab.edu

Abstract

Financial risk management increasingly depends on distributed data, heterogeneous models, and rapid coordination across institutions, yet the most valuable risk signals often remain trapped inside organizational boundaries because of privacy, competition, and regulatory constraints. This paper examines federated multi-agent learning as a system-level architecture for privacy-preserving financial risk management. Rather than treating privacy as a post hoc compliance layer, the discussion positions federated coordination, multi-agent decision intelligence, and distributed governance as interdependent design commitments. The paper analyzes architectural foundations, including client heterogeneity, communication efficiency, secure aggregation, and trust calibration across autonomous agents. It then considers privacy and security risks such as inference attacks, poisoning, model leakage, and adversarial coordination, arguing that differential privacy and secure computation must be paired with institutional accountability and continuous monitoring. Governance receives sustained attention because cross-border financial supervision, auditability, liability allocation, and fairness cannot be resolved through technical protocols alone. The paper also evaluates deployment constraints, socio-technical integration, systemic risk, scalability, and sustainability. It suggests that federated multi-agent learning can improve risk detection and resilience when it is embedded in layered infrastructures that combine technical safeguards, regulatory oversight, participatory governance, and human sensemaking. However, the same architecture can amplify opacity, herding, exclusion, and supervisory blind spots if incentives and accountability are misaligned. The conclusion proposes research and policy directions for robust, fair, and sustainable federated financial intelligence.

Keywords

federated learning, multi-agent systems, financial risk management, privacy preservation, governance, systemic risk, differential privacy, socio-technical systems.

1. Introduction

Financial risk management has become a data-intensive and computationally distributed enterprise. Banks, insurers, asset managers, payment networks, and regulators each hold partial views of exposure, liquidity, counterparty behavior, and market sentiment. Traditional centralized analytics can pool these views, but doing so often conflicts with privacy law, competitive secrecy, fiduciary duties, and operational security. Federated learning emerged as a response to this tension by enabling model training across decentralized data sources without requiring raw data centralization [1]. Subsequent research has expanded the approach into heterogeneous, non-independent, and institutionally asymmetric settings [2], [3]. In finance, such heterogeneity is not an edge case but the normal condition. Institutions differ in size, business model, regulatory jurisdiction, data quality, and risk appetite. Federated multi-agent learning therefore offers a conceptual bridge between privacy-preserving machine learning and distributed decision-making in financial infrastructures [4].

The multi-agent dimension is important because financial risk is not produced by a single model or institution. It emerges from interacting portfolios, trading strategies, credit exposures, liquidity channels, and supervisory responses. Multi-agent systems provide a vocabulary for autonomous or semi-autonomous entities that perceive their environment, exchange information, and adapt their behavior. When combined with federated learning, these agents can improve shared risk models while retaining local data and local decision authority. This architecture is attractive for credit scoring, fraud detection, anti-money-laundering analytics, market surveillance, operational risk, and climate-related financial stress testing. Yet the promise depends on whether privacy, robustness, fairness, and governance are treated as first-class system properties rather than secondary constraints.

The literature on federated learning has matured considerably. Foundational work demonstrated that decentralized training could approximate centralized performance under favorable conditions [1]. Later surveys identified open problems in communication, personalization, fairness, and adversarial resilience [2], [3]. System-oriented research examined production requirements such as scalability, fault tolerance, and secure aggregation [5]. Privacy research separately established that decentralized training alone does not guarantee confidentiality, because gradients, updates, and model parameters can leak sensitive information [6]. Differential privacy and secure computation provide stronger guarantees, but they introduce utility, cost, and coordination trade-offs [7], [8]. These technical findings are necessary but insufficient for financial risk management, where governance and institutional trust determine whether a model can be deployed at all [12].

Financial risk management also has distinctive temporal and systemic characteristics. Market regimes shift, correlations change, liquidity evaporates, and historical relationships break down. A model that performs well in stable periods may fail during crises. Multi-agent learning can adapt faster than centralized periodic retraining, but it can also synchronize behavior across institutions and create new channels for contagion. Recent work on trust-calibrated multi-agent decision intelligence addresses this problem by emphasizing reliability assessment and regime-sensitive coordination [14]. The implication is that federated financial intelligence must be evaluated not only by predictive accuracy but also by its effect on stability, contestability, and supervisory visibility.

This paper argues that federated multi-agent learning for privacy-preserving financial risk management should be understood as a socio-technical infrastructure. Its performance depends on model architecture, communication protocols, privacy budgets, incentive design, legal interoperability, and human interpretation. The sections that follow examine architectural foundations, privacy and security, governance, deployment, fairness and systemic risk, sustainability, and policy implications. The objective is to outline both the analytical promise and the structural hazards of distributing financial intelligence across autonomous agents.

2. Architectural Foundations of Federated Multi-Agent Learning

A federated multi-agent architecture for financial risk management typically consists of local agents, a coordination layer, shared model repositories, secure communication channels, and supervisory interfaces. Local agents retain data within institutional boundaries while contributing updates, signals, or policy parameters to a shared learning process. The coordination layer may be centralized, decentralized, or hybrid, depending on regulatory constraints and trust relationships. Centralized coordination can simplify aggregation and version control, but it creates a focal point for failure and control. Decentralized coordination reduces single-point dependency but complicates convergence, auditing, and accountability [2], [4]. Hybrid designs are common in finance because they allow a neutral utility or regulator to coordinate certain functions while leaving sensitive decisions local.

Heterogeneity is a defining architectural challenge. Financial institutions operate with different data schemas, reporting calendars, risk taxonomies, and model governance processes. Federated learning research often assumes comparable model architectures and similar data distributions, but financial data are non-identically distributed across institutions and time [3]. A large bank may have dense transaction data, while a regional insurer may have sparse but specialized claims data. A payment network may observe cross-border flows that individual banks cannot see. Multi-agent learning must therefore support personalization, meta-learning, or modular aggregation rather than forcing a single global model onto all participants. Otherwise, the shared model may improve aggregate performance while degrading local risk control.

Communication efficiency is another structural concern. Financial risk models may need to update frequently during volatile periods, but communication costs, latency, and bandwidth constraints can limit real-time coordination. Compression, sparsification, asynchronous updates, and hierarchical aggregation have been proposed to reduce communication overhead [11]. In multi-agent settings, communication is not merely a technical cost; it is also a governance event. Every update exchanged between agents can reveal information, trigger compliance obligations, or alter competitive dynamics. Architectural choices about what is shared, when it is shared, and who can observe it therefore have legal and strategic consequences.

Secure aggregation and trusted execution environments are often used to protect model updates during federation. Secure aggregation allows a server to compute an aggregate without seeing individual contributions, while trusted execution can isolate sensitive computations. These mechanisms are valuable, but they do not eliminate all risks. They may increase computational overhead, complicate debugging, and create new trust assumptions about hardware vendors or coordinating entities [5]. In finance, where operational resilience and supervisory access are critical, such assumptions must be explicit. A system that is private against other participants but opaque to regulators may be unacceptable. Conversely, a

system that is fully transparent to regulators may expose proprietary information to competitors if access controls fail.

Multi-agent coordination introduces questions of goal alignment and decision authority. Agents may optimize local risk-adjusted returns, regulatory capital, fraud loss, customer fairness, or liquidity buffers. These objectives can conflict. A federated fraud detection agent may benefit from sharing suspicious patterns, while a business unit may resist sharing because it fears reputational damage or customer attrition. Coordination mechanisms must therefore incorporate negotiation, incentive compatibility, and escalation pathways. The system should not assume that all agents share a single objective function. Instead, it should make trade-offs visible and governable.

Finally, architectural foundations must include observability and auditability. Financial risk systems are subject to internal audit, external audit, regulatory examination, and dispute resolution. A federated multi-agent system must be able to reconstruct why a decision was made, which updates influenced it, and whether privacy protections were applied correctly. This requires logging, provenance, model versioning, and access control that work across institutional boundaries [5], [10]. Without such observability, the architecture may be technically innovative but institutionally unusable.

3. Privacy, Trust, and Security in Financial Risk Infrastructures

Privacy is the primary motivation for federated approaches, but it is also the most easily misunderstood. Federated learning prevents raw data from being directly centralized; it does not automatically prevent information leakage. Gradients and model updates can encode sensitive attributes, transaction patterns, or client behavior. Membership inference, property inference, and reconstruction attacks have demonstrated that decentralized training can still expose private information [6]. In financial contexts, leakage could reveal trading positions, creditworthiness, fraud investigations, or politically exposed person relationships. Privacy must therefore be enforced through multiple layers, including differential privacy, secure aggregation, update clipping, and strict access control [7], [8].

Differential privacy provides a formal framework for limiting the influence of any individual or institution on published outputs. It is attractive because it offers quantifiable guarantees and composes across queries. However, it introduces a privacy-utility trade-off that is especially difficult in finance. Too much noise can obscure rare fraud patterns or tail risks; too little noise can leave sensitive information exposed. Setting privacy budgets requires not only technical judgment but also legal, ethical, and business judgment. Institutions may have different tolerances for privacy loss, and regulators may demand minimum protections. A federated system must therefore support heterogeneous privacy policies without collapsing into the lowest common denominator [8], [12].

Trust in federated multi-agent systems is not a binary property. It is calibrated across participants, updates, model versions, and decision contexts. An agent may trust another agent's aggregate contribution but not its individual updates. A regulator may trust a model's provenance but not its performance during a crisis. Trust calibration involves evaluating reputation, historical reliability, validation performance, and institutional accountability [14]. In financial risk management, trust must also be contestable. If an agent's contribution is downgraded, the system should provide a mechanism for appeal, audit, and remediation. Otherwise, trust scoring becomes a new form of opaque power.

Security threats extend beyond privacy leakage. Poisoning attacks can corrupt shared models by submitting malicious updates, while backdoor attacks can embed triggers that activate under specific market conditions. Sybil attacks can allow a single actor to simulate multiple participants and distort aggregation. In a multi-agent financial system, adversaries may include external criminals, compromised institutions, or insiders seeking competitive advantage. Defenses include robust aggregation, anomaly detection, reputation systems, and cryptographic identity. Yet these defenses can themselves be gamed. Robust aggregation may exclude legitimate minority institutions; reputation systems may entrench incumbents; identity systems may raise privacy concerns [2], [6].

Operational security is equally important. Federated systems depend on APIs, orchestration services, key management, and model registries. Each component can become an attack surface. A compromised coordination node could alter aggregation rules, leak updates, or deny service during a liquidity crisis. A compromised local agent could poison the global model while appearing compliant. Financial regulators increasingly expect operational resilience, incident reporting, and third-party risk management. Federated architectures must therefore be designed for graceful degradation, anomaly detection, and recovery. They should not assume that all participants remain honest or available.

Finally, privacy and security cannot be separated from institutional trust. Financial markets rely on confidence that sensitive information is handled responsibly and that systemic risks are visible to supervisors. A federated system that protects privacy but prevents regulators from detecting emerging contagion may undermine stability. A system that gives regulators broad access but exposes proprietary strategies may discourage participation. The design challenge is to create differentiated visibility: privacy for market participants, aggregate transparency for supervisors, and auditable accountability for the public. This requires legal and technical co-design rather than a purely cryptographic solution [12], [13].

4. Governance, Regulatory Compliance, and Institutional Design

Governance is the central bottleneck for federated multi-agent learning in finance. Technical systems can be built quickly, but legal mandates, supervisory expectations, and institutional incentives change slowly. The General Data Protection Regulation and analogous regimes impose conditions on data processing, cross-border transfer, purpose limitation, and individual rights [12]. Federated learning can support compliance by reducing raw data movement, but it does not eliminate the need for lawful basis, transparency, and data protection impact assessments. Model updates may still constitute personal data if they relate to identifiable individuals. Governance must therefore address the entire lifecycle of federated intelligence, from data selection to model retirement.

Ethical governance has moved from abstract principles to operational expectations. Frameworks for artificial intelligence ethics emphasize fairness, accountability, transparency, human oversight, and non-maleficence [13], [17]. In financial risk management, these principles translate into concrete requirements. A credit risk model must be explainable to affected customers. A fraud model must not disproportionately exclude vulnerable groups. A market surveillance model must preserve due process. A liquidity model must be stress-tested under adversarial conditions. Federated multi-agent systems complicate these requirements because responsibility is distributed across agents, coordinators, and data providers. Governance must assign accountability without pretending that a single actor controls the whole system.

Financial regulators have recognized the potential and risks of artificial intelligence in financial services [18]. Supervisory bodies are concerned with model risk management, outsourcing, data quality, cyber resilience, and systemic implications. Federated learning may align with regulatory goals by keeping data local and enabling privacy-preserving analytics. However, it also challenges supervisory models built around centralized institutions and discrete legal entities. If a shared model is trained across multiple banks, who is responsible for its errors? How can a supervisor examine a model that no single institution fully owns? How can a bank exit a federation without destabilizing the remaining participants? These questions require new governance instruments, including consortium agreements, model cards, audit rights, and supervisory protocols.

International coordination is especially difficult. Financial markets are global, but privacy and financial regulation remain jurisdictionally fragmented. A federated system spanning the European Union, United States, Singapore, and other regions must reconcile different data protection standards, secrecy laws, and supervisory powers. Bank for International Settlements guidance on fintech and supervisory practices highlights the need for cross-border cooperation and consistent risk management [19]. Federated learning could support regulatory technology that respects local rules while producing global risk insights. Yet without mutual recognition, safe harbor provisions, or multilateral supervisory agreements, participation may remain limited to jurisdictions with compatible legal frameworks.

Institutional design must also address incentives. Federated learning requires participants to contribute data, computation, and expertise. Why should a bank share updates that improve a competitor's model? Why should a smaller institution bear communication costs for a global model that favors larger players? Governance mechanisms must create fair value distribution, contribution measurement, and exit rights. Elinor Ostrom's work on governing commons suggests that durable institutions rely on clearly defined boundaries, collective choice, monitoring, graduated sanctions, and conflict resolution [22]. These principles apply to federated data and model commons. A federation is not just a network; it is a political economy of shared risk intelligence.

Finally, governance must incorporate sensemaking. Karl Weick's analysis of organizing emphasizes that actors interpret ambiguous signals and construct shared meaning through interaction [23]. In financial risk management, model outputs are not self-interpreting. A sudden anomaly may indicate fraud, cyberattack, market stress, or data error. Multi-agent systems can produce many signals, but human institutions must interpret them. Governance should therefore support deliberation, escalation, and learning rather than automated compliance. The goal is not to remove human judgment but to structure it with better evidence and clearer accountability.

5. Deployment Challenges and Socio-Technical Integration

Deploying federated multi-agent learning in financial institutions requires integration with existing systems, including core banking platforms, trading engines, risk engines, data warehouses, and regulatory reporting pipelines. These systems are often legacy, batch-oriented, and tightly coupled. Federated learning may require real-time or near-real-time APIs, event streams, and orchestration services. Production research on federated systems has shown that scalability, fault tolerance, and secure aggregation are difficult even in relatively controlled environments [5]. Financial deployment adds regulatory reporting, disaster recovery, and supervisory access. The result is a complex socio-technical program rather than a model deployment.

Latency and availability are critical in financial risk management. Fraud detection may need millisecond responses. Liquidity risk may need intraday updates. Credit risk may tolerate daily or weekly cycles. A federated multi-agent architecture must support differentiated service levels rather than a single synchrony model. Asynchronous updates can improve availability but complicate convergence and fairness. Hierarchical aggregation can reduce communication but introduces intermediate trust assumptions. Edge deployment can keep sensitive data local but limits computation. These trade-offs must be matched to risk domains and regulatory expectations [11].

Human-in-the-loop design is essential. Financial decisions have legal, ethical, and reputational consequences. A model may flag a transaction as fraudulent, deny a loan, or recommend a trading halt. In many cases, a human must review the decision or at least be able to contest it. Federated multi-agent systems can support human oversight by providing explanations, confidence scores, and provenance. However, explanations generated by one agent may not align with another's reasoning. Governance must define who can override, who must be notified, and how overrides are fed back into learning. Without such protocols, automation bias and accountability gaps may emerge.

Organizational culture is another deployment factor. Risk teams, data scientists, compliance officers, and business units often have different priorities and vocabularies. A federated system that produces sophisticated signals but does not fit existing workflows will be ignored or misused. Sensemaking theory suggests that new information technologies are adopted when they help actors reduce ambiguity and coordinate action [23]. Deployment should therefore include training, simulation, tabletop exercises, and crisis drills. Financial institutions already use stress testing and scenario analysis; these can be extended to federated multi-agent systems. A federation that has never practiced a coordinated response may fail when a real crisis occurs.

Cross-domain comparisons offer useful lessons. Healthcare federated learning faces similar privacy and governance challenges, but financial systems have stronger competitive secrecy and faster contagion dynamics. Supply chain federated learning emphasizes multi-party coordination, but financial risk has higher regulatory intensity. Smart grid federated learning deals with critical infrastructure, but financial markets add adversarial speculation and strategic behavior. These comparisons suggest that federated financial risk management requires a hybrid approach: privacy from healthcare, resilience from critical infrastructure, and market surveillance from finance. No single domain provides a complete template.

Finally, deployment must consider exit and failure. Institutions may leave a federation because of mergers, regulatory action, or strategic disagreement. A participant may be compromised and temporarily excluded. The coordination layer may fail during a market shock. The architecture should support graceful exit, data deletion, model rollback, and continuity of local risk management. These requirements are often overlooked in federated learning research, which focuses on training performance. In financial infrastructure, resilience is defined as much by failure handling as by normal operation [5], [19].

6. Fairness, Robustness, and Systemic Risk

Fairness in federated financial risk management has multiple meanings. At the individual level, models must not discriminate unlawfully or disproportionately harm protected groups. At the institutional level, smaller or specialized participants must not be systematically disadvantaged by aggregation rules. At the systemic level, the model must not create

correlated behavior that amplifies shocks. These levels interact. A model that improves average fraud detection but performs poorly for a regional bank may increase local risk. A model that optimizes credit approval for large banks may exclude thin-file customers. Fairness cannot be reduced to a single metric; it requires participatory evaluation and context-sensitive governance [13], [17].

Robustness is equally multidimensional. Financial models must withstand distributional shifts, adversarial manipulation, and structural breaks. Market regime shifts can invalidate historical correlations. A federated multi-agent system may adapt faster than a centralized model, but adaptation can also be destabilizing if agents react to the same signal simultaneously. Trust-calibrated coordination can help by weighting agents according to reliability and context, but trust itself can be manipulated [14]. Robustness therefore requires stress testing, adversarial simulation, and circuit breakers. A model that recommends rapid deleveraging across many institutions may trigger fire sales. Supervisory oversight must be able to detect and moderate such correlated responses.

Systemic risk is the most consequential financial concern. Federated learning can improve risk detection by combining distributed signals, but it can also create shared vulnerabilities. If many institutions rely on the same global model, a poisoned update or flawed parameter could propagate widely. If the model encourages similar trading or lending behavior, it may increase herding. If it obscures individual exposures, it may reduce market discipline. Financial stability authorities have long studied interconnection and contagion [18]. Federated multi-agent systems add a new layer of digital interconnection that must be mapped and monitored. The architecture should include systemic risk dashboards, scenario analysis, and supervisory access to aggregate exposures.

Financial machine learning research emphasizes the dangers of backtest overfitting and regime fragility [20]. A model that performs well in historical data may fail under new conditions. Federated learning can increase the amount of data available, but more data does not guarantee robustness if the data are non-representative or contaminated. Multi-agent systems can incorporate diverse perspectives, but diversity must be genuine rather than cosmetic. If all agents use similar features and objectives, federation may create an illusion of robustness while producing highly correlated errors. Governance should therefore require independent validation, out-of-sample testing, and adversarial review.

Investor sentiment and narrative dynamics also matter. Financial markets are shaped by collective interpretation, rumors, and media framing [21]. A federated risk system that ingests sentiment signals must distinguish noise from meaningful shifts. Multi-agent learning can model heterogeneous interpretations, but it can also amplify rumors if agents reinforce one another. Human oversight and supervisory communication are essential. During crises, public statements by regulators can stabilize or destabilize markets. A privacy-preserving federated system should not become an opaque source of authoritative risk signals without corresponding public accountability.

Finally, fairness and robustness must be integrated with privacy. A system that protects privacy but produces biased outcomes is not ethically acceptable. A system that is robust but Surveillance-intensive may violate legal norms. A system that is fair in aggregate but harmful locally may still fail. These tensions cannot be resolved through technical optimization alone. They require ongoing deliberation among institutions, regulators, civil society, and affected communities [13], [17]. Federated multi-agent learning should be designed to make these trade-offs visible and contestable.

7. Sustainability, Scalability, and Economic Viability

Sustainability in federated financial risk management has environmental, economic, and institutional dimensions. Training large models across many institutions consumes energy and computation. Secure aggregation and differential privacy add overhead. Frequent updates during volatile markets can increase network and processing costs. A sustainable architecture should use efficient communication, adaptive update schedules, and tiered model complexity. Not every risk domain requires real-time deep learning. Some problems may be addressed with simpler statistical models or periodic federated updates. Matching model complexity to decision value is both economically and environmentally prudent [2], [10].

Scalability is not only a matter of compute. It involves onboarding new participants, managing heterogeneous data schemas, resolving identity and access, and maintaining model quality as the federation grows. A small consortium of banks may coordinate through bilateral agreements. A large federation including insurers, payment providers, regulators, and fintech firms requires more formal governance. Scalability also affects fairness: as the number of participants grows, individual influence may diminish, and coordination costs may favor large institutions. Governance must include contribution measurement, weighting rules, and dispute resolution to prevent scale from becoming centralization [5], [22].

Economic viability depends on value creation and value capture. Participants must perceive tangible benefits, such as improved fraud detection, reduced compliance costs, better risk-adjusted pricing, or access to new markets. They must also bear costs, including integration, computation, privacy budget, and audit. A federated system is viable when benefits exceed costs for a sufficient number of participants and when the value is distributed fairly. Public goods such as financial stability may justify regulatory subsidies or mandates, but these must be designed carefully to avoid crowding out private innovation. Regulatory sandboxes and public-private pilots can help test economic models [19].

Institutional sustainability requires trust over time. A federation that is dominated by a few large players may be perceived as unfair. A coordination body that lacks legitimacy may struggle to enforce rules. A system that experiences a major privacy breach may lose participants permanently. Governance should therefore include transparency reports, independent audits, and mechanisms for redress. Ostrom's design principles for commons governance emphasize monitoring, graduated sanctions, and conflict resolution [22]. These principles can guide the creation of federated data and model commons in finance. They also suggest that sustainability is as much political as technical.

Finally, sustainability must consider model lifecycle. Models become outdated, data distributions shift, and regulations change. A federated system must support retraining, validation, retirement, and archival. Old model versions may be needed for audits or dispute resolution. Data deletion requests must be honored even if data contributed to past models. These requirements create tension between learning and forgetting. Technical mechanisms such as machine unlearning and privacy-preserving audit trails are emerging, but they remain immature. Financial institutions should plan for lifecycle governance from the beginning rather than treating it as an afterthought [12], [20].

8. Policy Implications and Future Research Directions

Policy makers face a difficult balance. They want to encourage innovation in financial risk management, protect consumers, preserve market stability, and maintain jurisdictional sovereignty. Federated multi-agent learning can support these goals if it is accompanied by

appropriate safeguards. Policy should promote interoperability, common standards, and supervisory access without mandating a single technical architecture. It should clarify liability for shared models, define acceptable privacy budgets, and require fairness testing. It should also support international cooperation, because financial risks do not stop at borders [18], [19].

Regulatory sandboxes can provide controlled environments for testing federated risk systems. They can allow institutions to experiment with secure aggregation, differential privacy, and multi-agent coordination under supervisory observation. Sandboxes should include civil society and consumer representatives, not only financial institutions and technology vendors. Ethical frameworks for artificial intelligence emphasize inclusiveness, accountability, and human oversight [13], [17]. These principles should be operationalized in sandbox criteria. For example, participants might be required to publish model cards, conduct disparate impact assessments, and provide redress mechanisms before deployment.

Future research should address several gaps. First, there is a need for evaluation metrics that capture privacy, fairness, robustness, and systemic risk together. Current benchmarks often isolate one dimension. Second, research should develop trust calibration methods that are resistant to manipulation and transparent to regulators. Third, work on federated multi-agent systems should incorporate legal and economic modeling, not only machine learning. Fourth, empirical studies are needed on how human decision-makers interpret multi-agent risk signals during crises. Fifth, research should examine exit, failure, and recovery in federated financial infrastructures [2], [5], [23].

Another research direction concerns the governance of model updates as speech or action. If an update influences credit access, market liquidity, or fraud investigations, it has social consequences. Governance must define when an update requires review, who can challenge it, and how harms are remedied. This is especially important for vulnerable populations and small institutions. A federated system should not become a private ordering regime that bypasses public law. Instead, it should be embedded in public accountability structures.

Finally, policy should anticipate strategic behavior. Institutions may join a federation to access shared intelligence while withholding their most valuable data. They may game contribution metrics or exploit trust scores. They may exit when convenient, leaving others exposed. Governance must include incentives for honest participation, penalties for free riding, and safeguards against collusion. These are classic problems in commons governance and multi-agent systems [22]. Addressing them requires interdisciplinary collaboration among computer scientists, economists, legal scholars, and financial supervisors.

9. Conclusion

Federated multi-agent learning offers a compelling architecture for privacy-preserving financial risk management. It allows institutions to learn from distributed data without centralizing sensitive records, and it supports adaptive coordination across heterogeneous agents. This paper has argued that the architecture should be understood as a socio-technical infrastructure. Its success depends not only on federated optimization and secure computation but also on governance, accountability, fairness, robustness, and sustainability. Privacy is necessary but not sufficient. Trust must be calibrated, contestable, and aligned with supervisory oversight.

The analysis has identified several structural trade-offs. Centralized coordination simplifies aggregation but creates control points. Decentralized coordination reduces dependency but complicates audit. Differential privacy protects individuals but can obscure rare risks. Multi-

agent adaptation improves responsiveness but can amplify herding. Scalability lowers barriers but may entrench large players. These trade-offs cannot be resolved once and for all. They require ongoing deliberation and adaptive governance.

Financial risk management is a domain where errors can harm individuals, institutions, and entire economies. Federated multi-agent learning should therefore be deployed with caution and evaluated against public interest criteria. It should enhance risk detection while preserving privacy, improve coordination while maintaining diversity, and support innovation while remaining accountable to regulators and affected communities. If designed well, it can contribute to a more resilient and inclusive financial system. If designed poorly, it may create new forms of opacity, exclusion, and systemic fragility. The path forward lies in interdisciplinary design that treats privacy, fairness, governance, and resilience as interdependent system properties.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

2. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

3. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

4. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.

5. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Van Overveldt, T., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374–388.

6. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1310–1321). ACM.

7. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). ACM.

8. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

9. Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., & Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

10. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775.

11. Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

12. Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer.

13. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

14. Liu, T. (2026). Trust-Calibrated Multi-Agent Decision Intelligence for Financial Risk Management under Market Regime Shifts. Available at SSRN 7307503.

15. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

16. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

17. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

18. Financial Stability Board. (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Financial Stability Board.

19. Bank for International Settlements. (2018). Sound practices: Implications of fintech developments for banks and bank supervisors. Bank for International Settlements.

20. Lopez de Prado, M. (2018). Advances in financial machine learning. Wiley.

21. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168.

22. Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge University Press.

23. Weick, K. E. (1995). Sensemaking in organizations. Sage.