Trust-Aware Multi-Agent Reinforcement Learning for Adaptive Financial Risk Management

Pleun Moenen
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
pleun.moenen164@ucf.edu

Enzio Gori
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
egori@uab.edu

Jasmijn van Saksen
Department of Computer Science, University of Houston, Houston, TX, USA.
vansaksen444@uh.edu

Candida Gordillo
Department of Computer Science, George Mason University, Fairfax, VA, USA.
candida1987@gmu.edu

Abstract

Financial risk management is shifting from centralized analytical pipelines toward distributed decision systems that combine machine learning, real-time market data, and autonomous or semi-autonomous agents. This paper examines trust-aware multi-agent reinforcement learning as an architectural and governance paradigm for adaptive financial risk management. The central argument is that trust cannot be treated as a secondary interface concern or as a simple confidence score. Instead, trust must be embedded as a system-level property that influences information sharing, action selection, escalation, auditability, and collective risk posture. The paper develops this argument through a system-oriented analysis of multi-agent learning architectures, market regime shifts, institutional governance, deployment infrastructure, sustainability, robustness, fairness, and policy implications. It considers how trust-aware coordination can improve resilience during stress events while also introducing new failure modes, including collusive convergence, opacity, correlated behavior, and accountability diffusion. The discussion emphasizes structural trade-offs between adaptability and stability, autonomy and control, efficiency and contestability, and local optimization and systemic prudence. It also connects financial risk management to broader socio-technical concerns such as algorithmic accountability, regulatory technology, and inclusive access to financial services. The paper concludes that trust-aware multi-agent reinforcement learning is most credible when designed as a governed infrastructure rather than as a purely predictive model. Its value depends on transparent incentives, contestable decision pathways, human oversight, stress-tested deployment, and continuous institutional learning across market cycles and regulatory regimes.

Keywords

trust-aware reinforcement learning; multi-agent systems; financial risk management; adaptive governance; systemic risk; market regime shifts; algorithmic accountability; socio-technical infrastructure.

1. Introduction

Financial risk management has become an information-intensive and computationally mediated field in which data volume, market velocity, and interconnected exposures exceed the capacity of purely manual oversight. Deep learning and reinforcement learning have expanded the technical repertoire available to financial institutions, offering methods for pattern recognition, sequential decision-making, and adaptive control under uncertainty [1]. Reinforcement learning, in particular, provides a framework for learning policies through interaction with an environment, which makes it attractive for dynamic portfolio adjustment, liquidity management, fraud detection, and stress response [2]. Traditional reinforcement learning texts emphasize the balance between exploration and exploitation, but financial environments add severe constraints: rewards are noisy, feedback is delayed, adversarial behavior is common, and mistakes can cascade across institutions [3]. Consequently, the question is not simply whether agents can learn profitable or defensive policies, but whether they can do so within governance structures that preserve trust, accountability, and systemic stability.

Multi-agent reinforcement learning extends these concerns by recognizing that financial markets are not single-agent environments. Banks, funds, exchanges, clearinghouses, regulators, and automated trading systems interact in cooperative, competitive, and sometimes ambiguous relationships [4]. Early Markov game formulations provided a foundation for understanding strategic interaction among multiple learners, while modern actor-critic methods have enabled scalable coordination in mixed cooperative-competitive settings [5]. Such approaches can model liquidity provision, order execution, risk transfer, and supervisory signaling as interdependent decisions rather than isolated optimization problems. Yet multi-agent learning also creates emergent dynamics that no single institution fully controls. Agents may learn to coordinate in benign ways, but they may also converge on correlated strategies that amplify volatility, exploit regulatory gaps, or externalize risk onto less powerful participants [6]. Trust therefore becomes a central design variable because it determines which signals are believed, which actions are authorized, which exceptions are escalated, and which actors are held accountable.

This paper argues for trust-aware multi-agent reinforcement learning as a system-level paradigm for adaptive financial risk management. Trust is understood here not as a sentiment or a static reputation score, but as a calibrated and contestable relationship among agents, institutions, infrastructures, and human supervisors. A trust-aware system must decide when to share information, when to defer, when to challenge, when to slow down, and when to escalate to human authority. It must also remain robust under market regime shifts, when historical correlations weaken and previously reliable policies become hazardous. The discussion is organized around architecture, governance, deployment, sustainability, fairness, robustness, and policy. The paper does not propose a single algorithm or a fixed model. Instead, it examines the structural conditions under which trust-aware multi-agent reinforcement learning can support financial risk management without intensifying opacity, herding, or accountability diffusion.

2. Trust as a System-Level Property in Financial Multi-Agent Systems

In financial systems, trust is often treated as a matter of institutional reputation, counterparty confidence, or regulatory assurance. A system-level view is broader. Trust is produced through technical verification, operational reliability, legal accountability, and shared expectations about behavior under stress. When multiple learning agents interact, trust must be represented in the architecture itself: through permissions, observability, redundancy, contestability, and escalation pathways. An agent that is highly accurate in normal conditions may still be untrustworthy if its reasoning cannot be inspected during a crisis or if its incentives reward short-term gains while externalizing long-term risk. This distinction matters because financial risk management is not only a prediction problem. It is a coordination problem in which the cost of misplaced confidence can be systemic.

Research on cooperation in social dilemmas shows that agents can learn prosocial or inequity-averse strategies when reward structures and interaction histories support them [7]. This finding has direct relevance to financial risk. If agents are rewarded only for individual profit or local loss reduction, they may learn to exploit shared resources such as market liquidity or supervisory attention. If they are rewarded for collective stability, they may accept short-term costs to preserve longer-term resilience. However, reward design alone cannot guarantee trustworthy behavior. Control-oriented perspectives on artificial intelligence emphasize that advanced systems require corrigibility, transparency, and bounded objectives [8]. Safety research further identifies concrete problems such as reward hacking, negative side effects, and scalable oversight, all of which can appear in financial settings when agents optimize proxies rather than genuine risk reduction [9]. Trust-aware design must therefore combine incentive alignment with architectural safeguards and institutional oversight.

Trust also has a temporal dimension. Financial trust is built through repeated interactions, successful recoveries, transparent post-mortems, and predictable responses to failure. An agent that performs well in calm markets may lose trust if it fails to explain anomalous behavior during a liquidity shock. Conversely, an agent that occasionally refuses to act because it detects uncertainty may preserve trust if that restraint is legible and justified. Risk management institutions already rely on conservative buffers, circuit breakers, and escalation thresholds, but trust-aware multi-agent learning requires these mechanisms to be adaptive rather than merely static. The system must learn not only when to take risk but also when to suspend learning, narrow autonomy, or transfer authority to human decision-makers. This implies that trust is both a computational state and an institutional relationship, and its governance cannot be reduced to a single metric.

3. Architecture for Trust-Aware Multi-Agent Reinforcement Learning

A credible architecture for trust-aware multi-agent reinforcement learning in finance must separate concerns that are often conflated in end-to-end learning systems. Perception, risk assessment, coordination, execution, audit, and human oversight should be distinguishable even when they are tightly integrated in operation. This separation supports contestability because supervisors can interrogate why a risk signal was produced, how it influenced a collective decision, and which agent or institution bore responsibility. Multi-agent actor-critic methods offer useful mechanisms for decentralized execution with centralized or shared learning signals [5]. Counterfactual credit assignment can help identify how individual contributions affect collective outcomes, which is important when many agents act simultaneously [6]. Yet financial applications require additional layers for permissioning, role differentiation, and regulatory reporting. A trading agent, a liquidity agent, and a compliance agent should not have identical objectives, information access, or autonomy.

Trust-aware coordination also depends on how information is represented and shared. In financial markets, data is fragmented across institutions, jurisdictions, and internal systems. Full transparency is neither feasible nor always desirable because it can reveal proprietary positions or create new attack surfaces. The architectural challenge is therefore selective disclosure: agents need enough shared context to assess correlated risk without exposing sensitive data to every participant. Trust mechanisms can mediate this exchange through attested data provenance, role-based access, privacy-preserving computation, and auditable query logs. These mechanisms do not eliminate the need for judgment, but they make trust more conditional and reviewable. A trust-aware system should be able to distinguish between confidence in a data source, confidence in a model, confidence in an agent's incentives, and confidence in the institution operating that agent. Treating these as a single dimension leads to brittle governance.

The literature on trust-calibrated decision intelligence for financial risk management under market regime shifts highlights the importance of aligning trust estimates with changes in market context [12]. When regimes shift, historical reliability can become misleading. An agent that was trustworthy in a low-volatility environment may become dangerous in a high-volatility or liquidity-constrained environment. A robust architecture must therefore include regime detection, uncertainty estimation, and dynamic autonomy. It must also include mechanisms for graceful degradation, such as reducing position sizes, increasing collateral buffers, routing decisions to human desks, or suspending certain strategies. Superintelligence and control debates remind us that autonomy without bounded authority can create catastrophic failure modes, even if the system is not generally superintelligent [13]. In finance, bounded authority means that agents operate within explicit mandates, loss limits, and escalation rules that cannot be overridden by short-term reward signals. Trust is thus encoded in the boundaries of action as much as in the accuracy of prediction.

4. Adaptive Risk Management under Regime Shifts

Financial risk management is fundamentally concerned with change. Volatility clusters, correlations shift, liquidity evaporates, and policy interventions alter the behavior of market participants. Empirical finance has long documented stylized facts such as fat tails, volatility clustering, and nonlinear dependence, which complicate the use of models calibrated on stable periods [11]. A risk system that performs well in one regime may fail in another because its learned representations encode relationships that no longer hold. Adaptive risk management therefore requires more than online learning. It requires mechanisms for detecting when adaptation is appropriate and when adaptation itself becomes a source of instability. If every agent rapidly updates its policy in response to the same market signal, the collective system may oscillate or herd, intensifying the very regime shift it is trying to survive. Trust-aware coordination can moderate this risk by introducing diversity, delayed response, and cross-checking among agents with different mandates.

Traditional risk management provides important safeguards, including capital adequacy, stress testing, limits, and governance committees [10]. These safeguards are not obsolete in an age of reinforcement learning; they become more important because learned policies can be opaque and because their failure modes may be correlated. A trust-aware multi-agent system should integrate conventional risk controls with adaptive learning rather than replacing one with the other. For example, an agent may propose a portfolio adjustment, but the adjustment should still pass through pre-trade risk checks, liquidity assessments, and regulatory constraints. During a regime shift, the system might reduce the influence of agents whose

recent performance is statistically indistinguishable from noise, or it might increase the weight of conservative agents with mandates to preserve capital. Such weighting is not merely a technical choice. It reflects an institutional judgment about which forms of expertise and caution deserve authority under uncertainty.

The supervisory dimension is equally important. Regulatory technology and supervisory technology have been promoted as ways to improve oversight, data collection, and real-time monitoring [18]. Trust-aware multi-agent reinforcement learning could support these goals by giving supervisors access to aggregated risk indicators, scenario simulations, and audit trails. However, it could also create new supervisory dependencies if agencies lack the expertise to interpret adaptive systems or if regulated institutions control the data pipelines on which oversight depends. The design challenge is to make adaptation legible without making it rigid. A system that cannot adapt will fail in novel markets, but a system that adapts too quickly or too opaquely will undermine accountability. The required balance is institutional as much as algorithmic: clear mandates, predefined triggers for escalation, independent validation, and periodic review of learned behavior across multiple market regimes.

5. Governance, Fairness, and Accountability

Governance is the mechanism through which technical trust becomes socially legitimate. Ethical frameworks for artificial intelligence emphasize human dignity, fairness, accountability, transparency, and beneficence, but these principles require operational translation in financial contexts [14]. A global survey of AI ethics guidelines shows convergence on high-level values alongside divergence in implementation and enforcement [15]. In financial risk management, governance must address who is accountable when a multi-agent system causes harm, how affected parties can seek redress, and how the benefits and burdens of automation are distributed. If trust-aware agents are deployed only by large institutions, they may improve resilience for those institutions while shifting risk to smaller counterparties, retail investors, or public balance sheets. Fairness therefore is not an external add-on. It is a condition for durable trust in the financial system.

Accountability in multi-agent settings is especially difficult because causation is distributed. An adverse outcome may emerge from interactions among agents that each behaved within their local mandates. Internal algorithmic auditing frameworks can help by documenting data provenance, model assumptions, testing procedures, and decision pathways [16]. But auditing must be ongoing rather than episodic, because multi-agent policies continue to learn and because market conditions change. Trust-aware systems should maintain immutable logs of significant actions, including the information available at the time, the confidence or uncertainty expressed by relevant agents, the escalation history, and the human approvals involved. These records support ex post accountability but can also support ex ante learning if they are used to revise mandates, retrain models, and redesign incentives. The goal is not to eliminate discretion but to make it traceable and contestable.

Fairness also intersects with market structure. If multiple institutions deploy similar reinforcement learning systems trained on similar data, their behaviors may become correlated even without explicit collusion. This can create systemic risk, reduce market diversity, and disadvantage participants who do not have access to comparable infrastructure. Governance must therefore consider concentration, interoperability, and access. Central banks and international bodies have emphasized the need for resilient digital infrastructures and effective supervisory tools [17]. Trust-aware multi-agent systems should be designed with public-interest considerations such as market integrity, financial inclusion, and operational

resilience. This may require regulatory sandboxes, shared benchmarks, disclosure requirements, and independent evaluation. It may also require limits on certain forms of autonomous coordination when they threaten fair competition or systemic stability. Trust without fairness is merely deference to power, and deference is a weak foundation for risk management.

6. Deployment, Infrastructure, and Sustainability

Deployment is where architectural ideals meet operational reality. Financial institutions operate under latency constraints, legacy system dependencies, regulatory reporting obligations, and cybersecurity threats. A trust-aware multi-agent reinforcement learning system cannot be deployed as a standalone model. It requires data pipelines, feature stores, model registries, simulation environments, approval workflows, monitoring dashboards, and incident response procedures. It also requires integration with existing risk engines, order management systems, and compliance platforms. These infrastructural demands are often underestimated because research prototypes abstract away from organizational complexity. In practice, the reliability of the system depends on the reliability of the surrounding infrastructure. If data arrives late, if model versions are inconsistent, or if escalation pathways are unclear, trust erodes quickly.

Sustainability has multiple meanings in this context. Environmental sustainability concerns the energy and resource costs of training and operating large learning systems, especially when many agents are continuously learning. Operational sustainability concerns whether the system can be maintained over time without excessive technical debt, vendor lock-in, or dependence on scarce expertise. Institutional sustainability concerns whether governance arrangements can survive leadership changes, market downturns, and regulatory scrutiny. A trust-aware architecture should therefore favor modularity, version control, reproducible experiments, and clear documentation. It should also include mechanisms for retiring obsolete agents and for preserving institutional memory. The history of high-frequency trading shows how technological change can reshape market microstructure, create new forms of advantage, and generate unintended consequences [19]. Trust-aware deployment should learn from these experiences rather than assume that greater automation automatically produces greater efficiency.

Infrastructure also shapes the distribution of trust. When a few firms control the most advanced models and the fastest execution environments, other participants must trust systems they cannot inspect. That asymmetry can be acceptable if oversight is strong, but it is fragile if oversight depends on the same firms for data and expertise. Public infrastructure such as secure data exchanges, common simulation platforms, and independent validation labs can help counterbalance private concentration. International coordination may be necessary because financial risk crosses borders and because regulatory arbitrage can undermine local safeguards [17]. The sustainability of trust-aware multi-agent reinforcement learning thus depends on a mixed ecology of private innovation, public oversight, and shared technical standards. No single institution can build trust alone, and no single regulator can monitor the entire system without cooperation.

7. Robustness, Stress, and Systemic Risk

Robustness in financial risk management means more than surviving a single adverse event. It means maintaining critical functions under stress, avoiding correlated failure, and recovering without amplifying the shock. Multi-agent reinforcement learning can improve robustness by

distributing decision-making and enabling local adaptation, but it can also reduce robustness by creating hidden dependencies and synchronized responses [2]. Deep reinforcement learning has achieved impressive performance in controlled environments, yet its policies can be brittle when the environment changes or when adversaries exploit learned weaknesses [20]. In financial markets, adversaries are persistent, incentives are strong, and the cost of failure is high. A trust-aware system must therefore be stress-tested against adversarial scenarios, extreme volatility, liquidity droughts, cyber incidents, and model manipulation. It must also be tested for emergent behavior that no individual agent was designed to produce.

Systemic risk arises when individual actions combine to produce collective harm. A trust-aware multi-agent system could mitigate this by sharing early warnings, coordinating circuit breakers, or reducing exposure when correlated behavior is detected. However, coordination itself can be dangerous if it becomes too tight or too opaque. If agents learn to coordinate in ways that resemble collusion, they may harm market integrity. If they coordinate too loosely, they may fail to prevent cascading losses. The governance challenge is to distinguish beneficial coordination from harmful coordination. This requires clear rules about information sharing, independent oversight, and legal boundaries. It also requires technical mechanisms for detecting abnormal synchronization and for injecting diversity when necessary. Trust-aware systems should not optimize for consensus alone. They should preserve the capacity for dissenting signals, especially when those signals come from agents with mandates to protect long-term stability.

Resilience also depends on human factors. Operators and risk managers must be able to understand, override, and learn from automated decisions. If trust-aware systems present recommendations as opaque outputs, humans may either over-rely on them or reject them entirely. Both responses are dangerous. Over-reliance leads to automation bias and slow response when the system fails. Rejection leads to inconsistent use and lost opportunities for improved risk management. The design should therefore support calibrated trust: confidence that is proportional to evidence, uncertainty, and context. This includes explanations at the right level of abstraction, uncertainty indicators, counterfactual scenarios, and clear escalation triggers. Human oversight is not a ceremonial layer. It is a functional component of the risk architecture, and its effectiveness must be measured and tested like any other control.

8. Policy Implications and Cross-Domain Comparisons

Policy for trust-aware multi-agent reinforcement learning must balance innovation with stability. Financial regulators already supervise model risk, operational risk, and market conduct, but multi-agent learning challenges conventional categories. It is difficult to classify a continuously learning system as a single model, a trading strategy, or an operational process. It may be all three. Regulatory frameworks may need to address the lifecycle of learning systems, including data governance, validation, change management, incident reporting, and decommissioning. They may also need to require evidence that trust mechanisms work under stress, not merely in backtests. Supervisory technology can assist, but it cannot substitute for legal accountability and institutional capacity [18]. Policy should encourage transparency without forcing disclosure that creates new vulnerabilities. It should promote common standards for audit logs, uncertainty reporting, and simulation environments while allowing room for responsible experimentation.

Cross-domain comparisons offer useful lessons. In autonomous vehicles, trust depends on safety certification, fail-safe behavior, and clear liability rules. In energy grids, trust depends on redundancy, decentralized control, and public oversight of critical infrastructure. In online

platforms, trust depends on content moderation, transparency reporting, and redress mechanisms. Financial risk management shares features with all these domains: high stakes, complex socio-technical systems, public interest, and asymmetric information. It also has distinctive features, including extreme speed, global interconnection, and the direct transfer of losses across institutions and households. Trust-aware multi-agent reinforcement learning should therefore borrow selectively from other domains while respecting finance's unique governance landscape. The goal is not to import a single regulatory model but to design adaptive institutions that can keep pace with evolving technical systems.

Fairness and inclusion are also policy concerns. If advanced risk management tools improve stability for large institutions but exclude smaller participants from safer infrastructure, the financial system may become more resilient in aggregate and less fair in distribution. Public policy can support shared utilities, open standards, and capacity-building for smaller firms and regulators. It can also require impact assessments that consider how automated risk decisions affect access to credit, insurance, and investment. Ethical frameworks emphasize beneficence and non-maleficence, but these principles require concrete metrics and enforcement [14]. Global guidelines provide a starting point, yet they vary in specificity and authority [15]. Trust-aware financial systems should be evaluated not only by their predictive accuracy but by their effects on market integrity, consumer protection, and the distribution of financial risk. Policy should make these effects visible and contestable.

9. Conclusion

Trust-aware multi-agent reinforcement learning offers a promising but demanding direction for adaptive financial risk management. It recognizes that modern financial systems are distributed, interactive, and continually changing, and that single-agent optimization is insufficient for understanding systemic outcomes. By embedding trust in architecture, governance, and deployment, institutions can coordinate more intelligently during regime shifts while preserving human oversight and accountability. However, the same properties that make multi-agent learning powerful also make it hazardous. Correlated adaptation, opaque coordination, reward hacking, and accountability diffusion can undermine stability and fairness if they are not governed carefully. Trust cannot be reduced to a confidence score or a reputation metric. It must be enacted through permissions, observability, escalation, redress, and continuous institutional learning.

The practical implication is that trust-aware systems should be treated as governed infrastructures rather than as isolated models. They require modular design, selective information sharing, robust stress testing, auditable logs, clear mandates, and mechanisms for graceful degradation. They also require policy frameworks that address lifecycle governance, market concentration, cross-border coordination, and distributional effects. The most credible deployments will be those that combine adaptive learning with conservative controls, that preserve diversity of behavior, and that make contestability a first-class design goal. Financial risk management ultimately depends on confidence among institutions, markets, and the public. If multi-agent reinforcement learning is to support that confidence, it must be designed to earn trust under uncertainty, not merely to optimize performance in favorable conditions.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

3. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

4. Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In Proceedings of the Eleventh International Conference on Machine Learning (pp. 157–163). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50027-1

5. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in Neural Information Processing Systems 30 (pp. 6379–6390).

6. Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2974–2982. https://doi.org/10.1609/aaai.v32i1.11794

7. Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., & Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In Advances in Neural Information Processing Systems 31 (pp. 3326–3336).

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

9. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

10. Hull, J. C. (2021). Risk management and financial institutions (6th ed.). Wiley.

11. Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. Quantitative Finance, 1(2), 223–236. https://doi.org/10.1080/713665670

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

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

14. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

15. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

16. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). https://doi.org/10.1145/3351095.3372873

17. Bank for International Settlements. (2020). Central banks and payments in the digital era. BIS Annual Economic Report. Bank for International Settlements.

18. Financial Stability Board. (2020). The use of supervisory and regulatory technology by authorities and regulated institutions. Financial Stability Board.

19. MacKenzie, D. (2018). Material signals: A historical sociology of high-frequency trading. American Journal of Sociology, 123(6), 1635–1683. https://doi.org/10.1086/697316

20. Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. Artificial Intelligence, 299, 103535. https://doi.org/10.1016/j.artint.2021.103535