ISSN: 3070-4014

Research Article

International Journal of Digital Journalism

# Trustworthy AI in Central Banking: International Practices and Financial Stability Risks

**Yue Dai***

*Postdoctoral Researcher Financial Research Institute, People's Bank of China*

***Corresponding Author**
Yue Dai, Postdoctoral Researcher Financial Research Institute, People's Bank of China.

**Submitted:** 2026, Jul 02 **Accepted:** 2026, Aug 07; **Published:** 2026, Aug 18

**Citation:** Dai, Y. (2026). Trustworthy AI in Central Banking: International Practices and Financial Stability Risks. *Int j Digital journalism, 2*(2), page no; 01-09.

***Abstract***

*Artificial intelligence (AI), and especially recent advances in large language models and foundation models, is becoming relevant to central banking in two distinct ways. It is a tool that can improve macroeconomic analysis, supervisory technology, payment oversight, anti-moneylaundering monitoring, cyber resilience, and financial stability surveillance. It is also a new source of risk that central banks must understand as stewards of monetary and financial stability. This paper reviews emerging international practices in the use of AI by central banks and financial authorities, and connects those practices to the computer science literature on trustworthy AI and AI safety. The central argument is that, for central banks, trustworthiness is not merely a model-level property. It is also a financial stability condition. Model failures, biased decision systems, privacy leakage, prompt injection, data poisoning, third-party concentration, algorithmic herding, and autonomous market responses can become systemic when many financial institutions rely on similar data, similar infrastructures, or similar foundation models. The paper proposes a framework that links trustworthy AI dimensions–robustness, safety, fairness, privacy, explainability, accountability, and human oversight–to core central bank functions. It concludes with a research agenda for central banks focused on model governance, AI-enabled supervision, privacy-preserving data cooperation, stress testing of AI-related shocks, and international coordination.*

**Keywords:** Central Banking, Trustworthy Ai, Artificial Intelligence, Financial Stability, Macroprudential Policy, Ai Safety, Supervisory Technology, Payment Systems

## 1. Introduction

Artificial intelligence has moved from a specialized research technology to a general-purpose infrastructure used across firms, households, and public institutions [1-3]. For central banks, this development matters for two reasons. First, AI is changing the economy that central banks must understand. It may affect productivity growth, capital expenditure, labor demand, wage dynamics, price-setting behavior, market structure, and the transmission of monetary policy. Second, AI is changing the tools available to central banks. Machine learning and large language models can help process unstructured data, extract signals from texts and transactions, detect anomalies, and support economic and financial analysis in real time [4-6]. Yet a central bank cannot treat AI as a neutral productivity tool. Finance is a high-stakes, highly interconnected, and confidence-sensitive system. A model that is merely unreliable in a low-risk consumer application may become systemically relevant when it informs credit allocation, liquidity management, trading, fraud detection, payment screening, stress testing, or supervisory monitoring. The question for central banks is therefore not only how AI can be used, but under what conditions AI can be trusted [7-9].

This paper reviews international central bank practices and frames them through the lens of trustworthy AI. The paper's main claim is simple: for central banks, trustworthy AI is not merely a property of individual models, it is an emerging condition for financial stability. Traditional discussions of trustworthy AI often focus on model-level attributes such as accuracy, robustness, fairness, privacy, explainability, and accountability. These attributes remain essential. But central banks must also consider system-

level properties: whether many financial institutions rely on the same vendor models, whether similar algorithms amplify market movements, whether AI agents accelerate liquidity hoarding or fire sales, whether synthetic data and model outputs pollute information environments, and whether cyber vulnerabilities in AI systems create operational risks for financial infrastructures.

The paper makes three contributions. First, it summarizes how central banks and financial authorities are approaching AI in monetary analysis, supervision, payment systems, anti-money laundering (AML), cyber security, and financial stability monitoring. Second, it connects central bank concerns to the technical literature on trustworthy AI and AI safety. Third, it proposes a research agenda that treats trustworthy AI as a bridge between computer science, financial regulation, and macroprudential policy.

### 1.1. Literature Landscape and Source Strategy

This review draws on two bodies of literature that have often developed separately. The first is the policy and central banking literature on AI adoption, digital finance, supervisory technology, financial stability, payment systems, anti-money-laundering, cyber resilience, and model risk [4,7,9-11]. The second body of literature comes from computer science and related fields, especially AI, data mining, security, privacy, and accountability venues. It provides concepts for foundation models, robustness, privacy, fairness, explainability, red-teaming, and AI safety [1,3,12-16]. The purpose of bringing these literatures together is not to claim that central bank AI governance can be reduced to technical AI safety, nor that technical AI safety can be reduced to financial regulation. Rather, the paper argues that central banks require a translation layer between the two. Concepts such as robustness, fairness, privacy, explainability, prompt injection, model extraction, data poisoning, and uncertainty calibration become macro-financially relevant when they are embedded in lending, payment, trading, supervision, compliance, and financial market infrastructure. Conversely, central banking concepts such as financial stability, market-wide liquidity, procyclicality, confidence, and operational resilience give trustworthy AI a system-level meaning that is rarely visible in model-level benchmarks.

### 1.2. AI as a Central Banking Issue

The Bank for International Settlements (BIS) has argued that AI affects central banks both as stewards of monetary and financial stability and as users of AI tools [4]. This dual role is useful for organizing the field. As stewards, central banks must understand how AI affects macroeconomic outcomes and financial system behavior. As users, they must decide how to incorporate AI into their own operations without weakening accountability, confidentiality, or policy judgment.

### 1.3. AI and The Macroeconomy

AI may affect both aggregate supply and aggregate demand. On the supply side, AI can raise productivity by augmenting cognitive work, improving prediction, reducing search and matching costs, automating routine tasks, and enabling new forms of innovation. On the demand side, AI may trigger large investment in data centers, chips, cloud infrastructure, software, and human capital. The net effect on inflation is ambiguous. Productivity improvements may lower inflationary pressure over time, while rapid investment demand and changes in market power may create shortrun price pressures [17-21].

Central banks must also study the labor market effects of AI. AI can generate new tasks and occupations, but it can also displace workers in activities that involve document processing, customer service, compliance screening, coding, and routine analysis. Distributional effects matter for monetary policy because they influence wages, consumption, credit demand, and political economy constraints around technological adoption [5,18,20]. Recent Federal Reserve communication illustrates this dual focus. Governor Lisa Cook described AI as relevant to both sides of the Federal Reserve's dual mandate–maximum employment and price stability–and emphasized that the Federal Reserve is studying AI's macroeconomic effects while using AI cautiously for internal research, coding, and writing support rather than for setting policy [5]. This distinction is important. AI may improve staff productivity and analytical breadth, but monetary policy remains a human, institutionally accountable judgment under uncertainty.

### 1.4. AI and Financial Intermediation

Financial institutions have long used statistical models, but modern AI changes the scale, scope, and autonomy of model use. In lending, AI can use alternative and unstructured data to improve default prediction and expand access for borrowers with thin credit files. In payments, AI can help detect fraud, identify suspicious transaction patterns, and reduce compliance costs. In asset management, AI can extract signals from news, filings, social media, order flow, and other high-frequency data. In insurance, AI can process claims, price risks, and analyze images or text [7,22-29]. These applications create efficiency gains, but they also create new forms of model risk. The same model can both improve credit scoring and encode hidden discrimination. The same anomaly detector can both reduce fraud and generate false positives that exclude legitimate users from payments. The same trading model can both improve price discovery and amplify market volatility if many institutions react to similar signals [30-33].

## 2. International Practices in Central Bank AI

International practices are still evolving. Central banks generally do not present AI as a replacement for policy judgment. Instead, they use AI as an analytical, supervisory, and operational tool, with strong attention to governance [5,6,8,30,31,34].

### 2.1. Macroeconomic Analysis and Nowcasting

Central banks increasingly use machine learning to process high-dimensional and unstructured data. Examples include using news, financial statements, firm earnings calls, surveys, job postings, online prices, payment data, electricity use, satellite data, and social media to nowcast economic activity and financial conditions. The BIS notes that large language models can transform unstructured information into structured signals and may improve nowcasting during periods when official data arrive with lags [4,35-38]. The

promise is not that AI replaces econometric reasoning. Rather, AI can expand the information set available to economists. Traditional macroeconomic models are often built for interpretability and structural reasoning, AI models are often built for prediction and pattern recognition. Central banks need hybrid approaches in which AI-generated signals are evaluated against economic theory, institutional knowledge, and real-time data constraints [39-43].

### 2.2. Supervision, AML, and Payment Oversight

Supervisory technology is one of the most immediate areas for AI adoption. Supervisors must review large volumes of regulatory filings, risk reports, transactions, complaints, news, and supervisory correspondence. AI can classify documents, summarize risk signals, detect anomalies, and prioritize supervisory attention [44-48]. AML and countering-the-financing-of-terrorism (CFT) monitoring are especially relevant. The BIS discusses Project Aurora, which used synthetic transaction data to compare machine learning methods for detecting suspicious money-laundering networks. Graph neural networks performed well because they could exploit payment relationships and network structure, and performance improved when data were pooled across institutions or jurisdictions [4,49]. This example highlights both the opportunity and the governance challenge. AML detection benefits from richer data sharing, but richer data sharing raises privacy, confidentiality, and cross-border legal issues [50-52].

### 2.3. Cyber Resilience

AI can strengthen cyber security by improving threat detection, automating routine security tasks, analyzing anomalies, and supporting red-team exercises. However, generative AI also strengthens attackers. It can produce more convincing phishing messages, imitate voice or writing styles, generate malware, or help attackers adapt their techniques. AI therefore changes both sides of the cyber security problem [53-58]. For central banks, cyber resilience is not merely an internal IT concern. Payment systems, settlement systems, central securities depositories, and other financial market infrastructures are critical infrastructure. AI-enabled attacks on these systems could create operational disruption, confidence shocks, and liquidity stress [60-62].

### 2.4. Internal Productivity and Controlled Experimentation

Many central banks are experimenting with AI for internal productivity: drafting, summarization, translation, code assistance, data cleaning, literature review, and knowledge management. The Federal Reserve's public discussion stresses responsible adoption, staff training, strong governance, privacy protection, cyber security, and human-in-the-loop oversight [5]. These principles are generalizable. Central banks should encourage learning and experimentation, but sensitive information, policy decisions, and externally visible actions require stricter controls [9,10,63-66].

## 3. Financial Stability Risks

AI-related financial stability risks arise when model-level failures interact with system-wide adoption. This section identifies six channels [7,8,19,32,34,67].

### 3.1. Model Opacity and Weak Explainability

Many AI systems are difficult to explain. This is a problem for financial firms, supervisors, and the public. In credit, insurance, and compliance, affected individuals and firms may need to know why a decision was made. In supervision, authorities need to understand why a model flags one institution as risky and not another. In macroprudential policy, explainability matters because policy actions require institutional legitimacy [68-71]. Explainability is not a binary property. Some models can be globally interpretable, others can be locally explained, still others can only be audited through behavioral testing. Central banks need to distinguish between low-risk uses where post-hoc explanations may be acceptable and high-risk uses where interpretability, auditability, and contestability should be required [63,72-74].

### 3.2. Bias, Discrimination, and Financial Inclusion

AI may expand financial inclusion by identifying creditworthy borrowers who are underserved by traditional credit scores. But it can also infer protected characteristics from proxies and reproduce historical discrimination. Bias can enter through training data, labels, sampling, feature selection, objective functions, and deployment feedback loops [14,75-81]. A model that optimizes default prediction without fairness constraints may reduce measured losses while worsening unequal access to credit or insurance.

For central banks and financial regulators, fairness is not only an ethical concern. It can affect trust in the financial system. If AI systems are perceived as arbitrary, discriminatory, or impossible to challenge, confidence in digital finance may decline [11,50,51,84,85].

### 3.3. Privacy Leakage and Data Governance

AI systems are data hungry. Financial data are sensitive, granular, and often legally protected. Payment data, account data, credit histories, transaction narratives, and supervisory reports can reveal personal behavior, business strategies, and institutional vulnerabilities. AI creates risks of data leakage through model training, memorization, prompt interfaces, logs, third-party APIs, and insecure integrations [86-89].

Privacy-preserving techniques such as differential privacy, federated learning, secure aggregation, and privacy-enhancing data clean rooms may help reconcile data cooperation with confidentiality [16,90-96]. However, these techniques are not automatic solutions. They involve trade-offs among statistical accuracy, privacy guarantees, implementation complexity, governance, and legal accountability [97-99].

### 3.4. Cyber-Attacks on AI Systems

AI systems create new attack surfaces. Prompt injection can manipulate model behavior. Data poisoning can corrupt training data. Model poisoning can compromise model integrity. Adversarial examples can induce wrong classifications. Model extraction and membership inference can reveal confidential information. The BIS explicitly identifies prompt injection, data

poisoning, and model poisoning as new cyber risks for financial institutions [4,42,100-105].

These risks are particularly important when AI systems are connected to tools: databases, payment workflows, compliance systems, customer communications, or trading platforms. A chatbot that merely answers questions is one risk category, an AI agent that can initiate actions is another [106-108].

### 3.5. Third-Party Concentration and Foundation Model Dependence

Cutting-edge AI requires data, compute, specialized talent, and capital. As a result, many financial institutions may rely on a small number of cloud providers, AI labs, model vendors, and data vendors. The BIS notes that the concentration of advanced AI provision creates third-party dependency risks for financial institutions [4,7,109,110]. A failure, outage, cyber-attack, model update, or governance error at a major provider could affect many institutions simultaneously. This is not merely traditional outsourcing risk. Foundation models can become shared cognitive infrastructure. If many firms use similar models for risk assessment, customer interaction, compliance, or trading, model behavior can become correlated across the financial system [2,3,111,112].

### 3.6. Algorithmic Herding and Procyclicality

The most distinctively macroprudential risk is that AI may amplify common behavior. Similar algorithms trained on similar data may respond to stress in similar ways. In normal times, this can look like efficiency. In stress, it can become herding, liquidity hoarding, fire sales, margin spirals, or abrupt credit tightening. AI can also accelerate decision speed beyond the capacity of human governance and supervisory intervention [7,8,32,34]. This channel connects model trustworthiness to financial stability. A model may be accurate on average and still be dangerous if it induces correlated actions under stress. Central banks therefore need stress tests that examine not only individual model performance but also the collective dynamics of AI adoption [20,33,113-115].

## 4. A Trustworthy AI Framework for Central Banking

The trustworthy AI literature offers useful concepts for central banks, but those concepts must be adapted to financial stability. Table 1 maps technical dimensions to central bank concerns [9,64,66,84,116]. The NIST AI Risk Management Framework defines AI risk management as a process for incorporating trustworthiness considerations into the design, development, use, and evaluation of AI systems [9,64]. For central banks, this logic should be extended from organizational risk management to system-wide financial risk management. A trustworthy AI model in finance should be evaluated not only by its local performance but also by its contribution to system resilience [63,117].

### 4.1. From Model-Level Trustworthiness to System-Level Trustworthiness

The computer science literature often begins with the model. Is it accurate? Is it robust? Is it fair? Is it explainable? Is it privacy-preserving? Central banks must begin with the system.

| Dimension | Technical question | Central bank relevance |
|---|---|---|
| Robustness | Does the model perform under distribution shift, adversarial inputs, and stress scenarios? | Crisis forecasting, market surveillance, payment anomaly detection, supervisory early warning. |
| Safety | Can the model or AI agent cause unintended harmful actions? | Tool-connected AI in payments, supervision, compliance, trading, and internal operations. |
| Fairness | Are errors and outcomes distributed unequally across groups? | Credit access, insurance pricing, consumer protection, financial inclusion. |
| Privacy | Can sensitive data be inferred, leaked, memorized, or misused? | Payment data, supervisory data, credit data, cross-border AML cooperation. |
| Explainability | Can decisions be understood, audited, and contested? | Supervisory accountability, model validation, consumer recourse, policy legitimacy. |
| Accountability | Who is responsible for model design, deployment, monitoring, and failure? | Governance of financial institutions, third-party vendors, and central bank AI use. |
| Systemic alignment | Do many AI systems create correlated behavior or destabilizing incentives? | Algorithmic herding, procyclicality, vendor concentration, macroprudential stress testing. |

**Table 1: Trustworthy AI Dimensions and Central Bank Relevance**

What happens if a class of models is adopted widely? What happens if an upstream data source is corrupted? What happens if a foundation model update changes behavior across many institutions at once? What happens if AI-generated market narratives influence both human traders and algorithmic systems [118-122]. This shift from model-level to system-level analysis is the key conceptual bridge between trustworthy AI and central banking. It suggests that central banks need model risk analysis, but also concentration analysis, network analysis, behavioral simulation, agent-based stress testing, and operational resilience assessment [59,123-125].

### 4.2. Human Oversight and Policy Judgment

Human oversight is a recurring principle in AI governance, but it requires careful specification. A formal human approval step is not sufficient if the human cannot understand the model, challenge

its output, or slow down automated processes. Human-in-the-loop governance should include clear responsibility, escalation procedures, audit trails, model documentation, training, and authority to halt AI systems that fail standards [12, 65,85,116,126]. In central banking, this point is especially important for monetary policy. AI can help summarize data, produce alternative scenarios, flag anomalies, and support research. It should not become a substitute for deliberation, institutional accountability, or judgment under uncertainty [4,5].

## 5. Research Agenda for Central Banks

The field now needs a research agenda that is both technical and institutional. Five priorities stand out.

### 5.1. AI Model Governance in Regulated Financial Institutions

Central banks and supervisors should develop methods for evaluating AI model governance across banks, insurers, payment firms, asset managers, and financial market infrastructures. Key questions include: How should firms document foundation-model use? How should they monitor model drift? What tests are required for fairness, robustness, privacy, and security? How should vendor models be audited when source code and training data are unavailable? Existing work on model cards, datasheets, risk management frameworks, and financial-sector AI principles provides a starting point, but it must be adapted to supervisory evidence and macroprudential use [9,11,52,63,72,122].

### 5.2. AI Stress Testing and Macroprudential Simulation

Traditional stress tests focus on balance sheets, capital, liquidity, and macroeconomic shocks. AI creates the need for new scenarios: a major model provider outage, a poisoned data feed, correlated AI-driven credit tightening, AI-generated misinformation affecting deposits or markets, autonomous trading models responding to the same signal, or a payment-system disruption caused by AI-enabled cyber-attacks. Central banks should explore agent-based simulations and network stress tests that capture correlated AI behavior [8,32,34,54,67].

### 5.3. Privacy-Preserving Cooperation

Many high-value AI use cases require cooperation across institutions and borders. AML, fraud detection, cyber threat intelligence, and payment monitoring all benefit from pooled information. But central banks must preserve privacy, confidentiality, and legal constraints. Research is needed on differential privacy, federated learning, secure aggregation, synthetic data, and privacy-preserving record linkage in real supervisory and payment contexts [16,49,80,92,94,96].

### 5.4. Evaluation Standards for Central Bank Ai Use

Central banks should hold their own AI use to high standards. Internal AI systems should be classified by risk. Low-risk productivity tools may require basic controls. Medium-risk analytical tools may require validation, documentation, and reproducibility checks. High-risk tools that affect supervision, market operations, or payment oversight should require strong human oversight, audit logs, red-team testing, and clear accountability [57,58,127-130].

### 5.5. International Coordination

AI risks cross borders. Foundation models, cloud infrastructure, payment networks, cyber threats, and financial markets are global. The BIS has called for central banks to build a community of practice for sharing knowledge, data, best practices, and AI tools [4]. This agenda should be expanded to include joint evaluation benchmarks, incident reporting taxonomies, cross-border AML experimentation, cyber exercises, and common principles for AI use in critical financial infrastructure [22,54,59,65,66,117].

## 6. Conclusion

AI is becoming part of the financial system's cognitive infrastructure. It helps institutions see, predict, classify, communicate, and decide. For central banks, this creates a double mandate of understanding and governance. They must understand how AI changes macroeconomic dynamics and financial intermediation. They must also govern the risks that arise when AI systems become embedded in payment systems, supervisory processes, trading, lending, insurance, cyber security, and institutional decision-making. The central message of this paper is that trustworthy AI in central banking must be defined at two levels. At the model level, AI should be robust, safe, fair, private, explainable, and accountable. At the system level, AI should not create hidden concentration, correlated behavior, operational fragility, or destabilizing feedback loops. A model can be locally useful and globally risky. That is why central banks need to connect AI safety research with financial stability analysis. The next stage of research should move from principles to measurement. Central banks need practical tests for AI robustness under stress, benchmarks for financial AI explainability, privacypreserving mechanisms for supervisory cooperation, red-team protocols for AI-enabled financial infrastructure, and macroprudential models of algorithmic herding. If central banks can build these capabilities, AI can become not only a source of risk, but also a tool for making the financial system more resilient.

## References


1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.
3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*
4. Bank for International Settlements (2024a). Artificial intelligence and the economy: Implications for central banks. BIS Annual Economic Report 2024, Chapter III.
5. Cook, L. D. (2025). AI: A Fed Policymaker's View. *Speech, Board of Governors of the Federal Reserve System*.
6. International Monetary Fund (2024a). Generative artificial intelligence and the financial sector.
7. Financial Stability Board (2017). Artificial intelligence and machine learning in financial services: Market developments

and financial stability implications.
8. Breeden, S. (2024). Engaging with the machine: Ai and financial stability. Speech, Bank of England.
9. National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (ai rmf 1.0).
10. Bank of England and Financial Conduct Authority (2022a). Artificial intelligence and machine learning. Discussion Paper DP5/22.
11. Monetary Authority of Singapore (2018). Principles to promote fairness, ethics, accountability and transparency in the use of artificial intelligence and data analytics in singapore's financial sector.
12. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
13. Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR), 54*(6), 1-35.
15. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence, 1*(5), 206-215.
16. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*.
17. Bank for International Settlements (2024b). Bis annual economic report 2024.
18. Cazzaniga, M., Jaumotte, M. F., Li, L., Melina, M. G., Panton, A. J., Pizzinelli, C., ... & Tavares, M. M. M. (2024). *Gen-AI: Artificial intelligence and the future of work*. International Monetary Fund.
19. International Monetary Fund (2024b). Global financial stability report.
20. Federal Reserve Bank of San Francisco (2024). Artificial intelligence, productivity, and the labor market.
21. Reserve Bank of Australia (2024). Artificial intelligence and the economy.
22. International Organization of Securities Commissions (2021). The use of artificial intelligence and machine learning by market intermediaries and asset managers.
23. Bank of England and Financial Conduct Authority (2024). Artificial intelligence and machine learning in UK financial services.
24. Bank of England and Financial Conduct Authority (2022b). Machine learning in uk financial services.
25. Bank of England and Financial Conduct Authority (2019). Machine learning in uk financial services.
26. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223-2273.
27. Berg, T., Burg, V., Gombović, A., & Puri, M. (2019). On the rise of fintechs–credit scoring using digital footprints. *Michael J. Brennan Irish Finance Working Paper Series Research Paper*, (18-12).
28. Jagtiani, J., & Lemieux, C. (2018). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform.
29. Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218-239.
30. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5-47.
31. Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review, 34*(3), 618-634.
32. Danielsson, J., & Uthemann, A. (2023). On the use of artificial intelligence in financial regulations and the impact on financial stability. *arXiv preprint arXiv:2310.11293*.
33. Sirignano, J., & Cont, R. (2018). Universal features of price formation in financial markets: perspectives from deep learning. *Available at SSRN 3141294*.
34. European Central Bank (2024b). Financial stability review.
35. Irving Fisher Committee on Central Bank Statistics (2017). Big data. IFC Bulletin No. 44.
36. Irving Fisher Committee on Central Bank Statistics (2019). The use of big data analytics and artificial intelligence in central banking. IFC Bulletin.
37. Federal Reserve Bank of New York (2019). Nowcasting report and machine learning approaches to real-time economic measurement.
38. European Central Bank (2022). Text-based indicators and monetary policy analysis. ECB Working Paper Series.
39. Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
40. Yang, Y., Uy, M. C. S., and Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.
41. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
42. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., ... & Liu, Y. (2023). Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.
43. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474.
44. Broeders, D. and Prenio, J. (2018). Innovative technology in financial supervision: Suptech. FSI Insights on Policy Implementation No. 9.
45. di Castri, S., Grasser, M., and Kulenkampff, A. (2019). The suptech generations. FSI Insights on Policy Implementation No. 19.
46. BIS Innovation Hub (2022). Project ellipse: Regulatory data and analytics platform.
47. European Central Bank (2023). Suptech tools for banking

supervision.
48. European Central Bank Banking Supervision (2024). Artificial intelligence and banking supervision.
49. BIS Innovation Hub (2023b). Project aurora: The power of data, technology and collaboration to combat money laundering.
50. Monetary Authority of Singapore (2022). Veritas initiative: Responsible ai in financial services.
51. Monetary Authority of Singapore (2023). Veritas toolkit for fairness, ethics, accountability and transparency
52. Hong Kong Monetary Authority (2019). High-level principles on artificial intelligence.
53. BIS Innovation Hub (2023c). Project raven: Cyber resilience in central banking.
54. Financial Stability Board (2023). Format for incident reporting exchange.
55. European Central Bank (2024a). Cyber resilience and digital operational resilience in the financial sector.
56. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228.*
57. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022, December). Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3419-3448).
58. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., ... & Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
59. Financial Stability Board (2024). Enhancing third-party risk management and oversight.
60. BIS Innovation Hub (2021). Project mbridge: Connecting economies through cbdc.
61. BIS Innovation Hub (2023a). Project atlas: Mapping the world of decentralised finance.
62. Bank for International Settlements (2022). The future monetary system. BIS Annual Economic Report 2022, Chapter III.
63. Bank of England Prudential Regulation Authority (2023). Model risk management principles for banks.
64. National Institute of Standards and Technology (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile.
65. The White House (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.
66. European Union (2024). Artificial intelligence act.
67. Board of Governors of the Federal Reserve System (2024). Financial stability report.
68. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
69. Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
70. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.
71. Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. *In International conference on machine learning* (pp. 1885-1894). PMLR.
72. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. I*n Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).
73. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.
74. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
75. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.
76. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
77. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153-163.
78. Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In 20th International Conference on Artificial Intelligence and Statistics (pp. 962-970).
79. Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems, 30*.
80. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).
81. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*, 671.
82. Monetary Authority of Singapore (2022). Veritas initiative: Responsible ai in financial services.
83. Monetary Authority of Singapore (2023). Veritas toolkit for fairness, ethics, accountability and transparency.
84. European Commission High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy ai.
85. Hong Kong Monetary Authority (2024). Consumer protection in respect of use of generative artificial intelligence.
86. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)* (pp. 2633-2650).
87. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.

88. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium* (*USENIX Security 16*) (pp. 601-618).
89. Song, C., Ristenpart, T., and Shmatikov, V. (2017). Machine learning models that remember too much. In *ACM Conference on Computer and Communications Security*.
90. Dwork, C., McSherry, F., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Lecture notes in computer science*.
91. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science, 9*(3-4), 211-487.
92. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308-318).
93. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). Pmlr.
94. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), 10(2), 1-19.
95. 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.
96. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017, October). Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175-1191).
97. Nasr, M., Shokri, R., & Houmansadr, A. (2019, May). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)* (pp. 739-753). IEEE.
98. Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017, October). Deep models under the GAN: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 603-618).
99. Jagielski, M., Ullman, J., & Oprea, A. (2020). Auditing differentially private machine learning: How private is private sgd?. *Advances in Neural Information Processing Systems, 33*, 22205-22216.
100. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.
101. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018, April). Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)* (pp. 399-414). IEEE.
102. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206*.6389.
103. Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems, 31*.
104. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.
105. Liu, Y., Ma, S., Aafer, Y., Lee, W. C., Zhai, J., Wang, W., & Zhang, X. (2018, January). Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium* (*NDSS 2018*). Internet Soc.
106. Yao, S., Zhao, J., Yu, D., Shafran, I., Narasimhan, K. R., & Cao, Y. (2022, October). React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.
107. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
108. Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., ... & Xiong, C. (2022). Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.
109. Crisanto, J. C., Ehrentraud, J., & Fabian, M. (2021). Big techs in finance: regulatory approaches and policy options.
110. Feyen, E., Frost, J., Gambacorta, L., Natarajan, H., & Saal, M. (2021). Fintech and the digital transformation of financial services: implications for market structure and public policy. *BIS papers*.
111. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Sharick, E. (2023). Paper Review:'Sparks of Artificial General Intelligence: Early experiments with GPT-4'.
112. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
113. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.
114. Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., ... & Nado, Z. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift.
115. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems, 30*.
116. OECD (2019). Oecd principles on artificial intelligence.
117. Monetary Authority of Singapore (2024). Generative ai risk and governance in financial services.
118. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477-486.

119. Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., ... & Virk, G. (2023, August). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 723-741).
120. Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*.
121. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
122. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
123. Auer, R. (2019). Embedded supervision: how to build regulation into blockchain finance. *Globalization and Monetary Policy Institute Working Paper*, (371).
124. Aramonte, S., Huang, W., and Schrimpf, A. (2021). Defi risks and the decentralisation illusion. BIS Quarterly Review.
125. Bank for International Settlements (2023). The financial stability risks of decentralised finance. BIS Bulletin.
126. Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Uk.
127. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
128. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
129. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology, 15*(3), 1-45.
130. Lin, S., Hilton, J., & Evans, O. (2022, May). Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 3214-3252).