

# Agentic AI for Automated Stress Testing and Financial Risk Scenario Analysis

Dawid Dorociak

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.  
hellodawid@uc.edu

Maryam Kirpenstein

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.  
maryamwork@colostate.edu

Orazio Udinesi

Department of Computer Science, University of Central Florida, Orlando, FL, USA.  
helloorazio@ucf.edu

Yashvi Chandran

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,  
OR, USA.  
yashvi108@oregonstate.edu

## Abstract

Agentic artificial intelligence is emerging as a transformative layer for automated stress testing and financial risk scenario analysis. Unlike conventional models that produce forecasts or classifications within narrowly specified tasks, agentic systems can plan, coordinate tools, retrieve evidence, simulate counterfactuals, and adapt their behavior across changing market regimes. This paper examines the system-level implications of deploying such systems in banks, regulators, insurers, asset managers, and market infrastructures. It argues that the value of agentic AI in stress testing depends less on isolated predictive accuracy than on the architecture that connects scenario generation, model orchestration, data provenance, human oversight, and regulatory accountability. The paper develops a conceptual framework for agentic stress testing as a socio-technical infrastructure, emphasizing structural trade-offs among autonomy, auditability, robustness, fairness, and sustainability. It discusses how multi-agent designs can improve coverage of tail events, climate transition pathways, cyber shocks, liquidity spirals, and sovereign-bank feedback loops, while also introducing risks of correlated behavior, model monoculture, opaque reasoning, and governance drift. The analysis highlights the need for trust calibration, adversarial validation, invariant learning, model risk controls, and continuous auditing. It further considers deployment constraints, computational cost, energy intensity, and institutional capacity. The conclusion proposes that agentic stress testing should be governed as critical financial infrastructure, with clear escalation paths, contestability mechanisms, and policy safeguards that preserve human responsibility under uncertainty.

## Keywords

agentic AI; financial risk; stress testing; scenario analysis; multi-agent systems; model governance; systemic risk; human oversight.

## 1. Introduction

Financial stress testing has become a central instrument of prudential regulation, institutional risk management, and systemic risk surveillance. Its purpose is not merely to estimate losses under adverse conditions but to expose vulnerabilities, inform capital and liquidity planning, and support policy decisions when uncertainty is high. Traditional stress testing frameworks combine macroeconomic scenarios, supervisory models, balance sheet projections, and expert judgment. They have proven valuable, yet they remain constrained by static scenario libraries, fragmented data pipelines, slow update cycles, and limited ability to represent adaptive market behavior. The growing complexity of climate risk, cyber shocks, nonbank intermediation, and cross-border spillovers intensifies the demand for more dynamic and exploratory approaches.

Agentic AI offers a different paradigm. Rather than treating scenario analysis as a sequence of manually curated assumptions and model runs, agentic systems can orchestrate heterogeneous tools, interrogate data sources, generate alternative narratives, test assumptions, and refine scenarios through iterative reasoning. Foundational work on artificial intelligence control emphasizes that increasingly capable systems require robust oversight mechanisms rather than simple performance optimization [1]. Multi-agent systems research provides concepts for coordination, negotiation, and distributed problem solving that are directly relevant to financial institutions composed of desks, risk functions, regulators, and market infrastructures [2]. Deep learning and reinforcement learning demonstrate how complex representations and sequential decision making can be learned from data and simulated environments [3][4]. These capabilities create opportunities for stress testing that is more adaptive, granular, and responsive to regime shifts.

However, the financial system is not a closed laboratory. It is a socio-technical infrastructure shaped by incentives, regulation, market microstructure, and human interpretation. Supervisory institutions have long stressed the importance of rigorous scenario design and transparent methodologies [5]. Climate-related financial stability analysis has shown that forward-looking risk assessment must confront deep uncertainty, long horizons, and non-linear tipping points [6]. Basel stress testing principles similarly emphasize governance, documentation, and use in decision making, not only model output [7]. Agentic AI must therefore be evaluated as part of a broader institutional architecture. If deployed without appropriate controls, it may amplify procyclicality, create false confidence, obscure accountability, or produce correlated reactions across institutions that intensify systemic stress. The central question is not whether agentic AI can generate scenarios, but under what governance, infrastructure, and epistemic conditions it should be allowed to influence financial decisions.

## **2. From Automation to Agency in Financial Risk Workflows**

Automation in finance has traditionally referred to rule-based processing, statistical scoring, or model-driven calculations. Agentic AI extends automation by giving systems degrees of goal-directed initiative. An agent may select tools, decompose tasks, request data, evaluate intermediate results, and decide when to escalate to a human. In stress testing, this shift is significant because scenario analysis is not a single prediction problem. It involves framing the shock, selecting transmission channels, calibrating severity, interpreting model disagreement, and communicating uncertainty. Recent work on trust-calibrated multi-agent decision intelligence for financial risk management under market regime shifts points toward architectures that explicitly balance agent autonomy with confidence estimation and regime

awareness [8]. Such approaches recognize that a system's willingness to act should depend on the reliability of its evidence and the stability of the environment.

The movement from automation to agency also changes labor and accountability. A system that merely calculates a risk metric can be validated through standard model risk management. A system that selects scenarios, prioritizes vulnerabilities, and recommends actions becomes a participant in decision making. This raises questions about responsibility, contestability, and the boundary between machine recommendation and human judgment. Research on automation and labor markets shows that technology can substitute for some tasks while complementing and reorganizing others [9]. In risk functions, the likely outcome is not full replacement of analysts but a redistribution of cognitive labor toward supervision, validation, and exception handling. Agents may handle high-volume exploratory analysis, while humans focus on framing assumptions, adjudicating conflicts, and taking responsibility for consequential choices.

A further structural issue is robustness under distribution shift. Financial regimes are not stationary. Interest rate environments, liquidity conditions, geopolitical alignments, and market microstructure can change abruptly. Methods such as invariant risk minimization attempt to learn representations that remain stable across environments rather than exploiting spurious correlations [10]. In an agentic stress testing system, invariant principles could inform how scenario generators are evaluated across historical regimes, synthetic perturbations, and adversarial conditions. However, no technical method eliminates the need for institutional judgment. The system must be able to recognize when its assumptions no longer hold and when it should defer to human authority. Agency without calibrated deference is dangerous in finance, where confidence can become a source of contagion.

### **3. Architecture of Agentic Stress Testing Systems**

An agentic stress testing architecture is best understood as a layered system rather than a single model. The data layer integrates supervisory returns, market data, internal exposures, climate indicators, cyber telemetry, and textual sources. The representation layer converts these inputs into entities, relationships, and temporal states. The agent layer contains specialized agents for scenario generation, transmission channel analysis, portfolio projection, liquidity simulation, and validation. The orchestration layer coordinates tasks, resolves conflicts, and enforces policy constraints. The governance layer records decisions, manages permissions, and provides audit trails. Each layer introduces trade-offs among flexibility, latency, cost, and control.

Specialized agents can improve coverage by allowing different perspectives to compete. A macroeconomic agent might model inflation and unemployment pathways; a market liquidity agent might simulate fire sales and margin spirals; a climate agent might translate physical and transition hazards into sectoral shocks; a cyber agent might represent operational disruptions and confidence effects. Multi-agent deliberation can surface disagreements that a single monolithic model would suppress. Yet specialization also creates integration problems. Agents may use incompatible ontologies, duplicate effort, or produce contradictory recommendations. Without a strong orchestration layer, the system can become a collection of plausible but incoherent narratives. This is analogous to hidden technical debt in machine learning systems, where data dependencies, configuration drift, and entanglement create maintenance burdens that are not visible in model accuracy metrics [11].

Safety and reliability must be designed into the architecture. Concrete problems in AI safety include reward hacking, negative side effects, scalable oversight, and safe exploration [12]. In stress testing, reward hacking might appear as agents optimizing for scenarios that produce dramatic but implausible losses, or conversely, selecting benign scenarios that minimize capital impact. Negative side effects could include excessive data access, computational cost, or market-moving information leakage. Scalable oversight becomes difficult when agents generate complex reasoning that humans cannot fully inspect in real time. The architecture must therefore include constraints, sandboxes, permissioning, provenance tracking, and mandatory review points. Agentic stress testing should not be a black box that outputs a number; it should be an inspectable process that exposes assumptions, evidence, and uncertainty.

#### **4. Scenario Generation and Regime-Shift Detection**

Scenario generation is the creative and epistemic core of stress testing. Traditional approaches rely on expert-designed scenarios, historical episodes, and parametric shocks. Agentic systems can augment this by generating narratives, combining shocks, and exploring tail dependencies. They can simulate the reactions of households, firms, banks, and policymakers, creating a richer representation of feedback loops. However, generative capacity is not the same as validity. A scenario is useful only if it is plausible, decision-relevant, and grounded in causal mechanisms. Agents must therefore be constrained by economic theory, institutional knowledge, and empirical regularities, while still being allowed to explore unconventional combinations.

Regime-shift detection is essential for automated scenario analysis. If an agent assumes that historical correlations will persist, it may underestimate the severity of a transition from low inflation to tightening, from stable funding to liquidity stress, or from gradual climate policy to abrupt repricing. Techniques for monitoring structural breaks, volatility regimes, and correlation instability can inform when the system should expand its scenario space. The trust-calibrated multi-agent approach cited earlier emphasizes that trust in agent decisions should be adjusted when market regimes shift [8]. This is a critical design principle: the system should not only produce scenarios but also estimate the reliability of its own operating assumptions. When confidence falls, the system should shift from autonomous exploration to human-supervised deliberation.

Scenario generation also has distributional and behavioral dimensions. Prospect theory shows that decision making under risk is shaped by loss aversion, framing, and reference dependence [16]. If agents model market participants as purely rational expected-utility maximizers, they may miss the behavioral amplification of stress. Similarly, black swan theory warns that rare, high-impact events are often outside standard models and may be rationalized only after the fact [17]. Agentic systems could help by maintaining a library of outlier narratives and testing institutional resilience against them. Yet they can also create an illusion of completeness if they generate many scenarios without acknowledging the limits of imagination. The governance challenge is to treat scenario generation as a disciplined form of organized skepticism rather than a production line for risk numbers.

#### **5. Governance, Auditability, and Accountability**

Governance is the decisive factor in whether agentic stress testing strengthens or destabilizes financial risk management. A system that can autonomously generate scenarios and recommend actions must be subject to clear authority boundaries. Who can modify agent

goals? Who approves new tools? Who interprets agent disagreement? Who is accountable when an agent-recommended scenario leads to inadequate capital planning? These questions cannot be resolved by technical accuracy alone. Research on algorithmic auditing argues for end-to-end accountability frameworks that span design, development, deployment, and monitoring [13]. In financial stress testing, such frameworks must align with existing model risk management, internal audit, supervisory review, and board oversight.

Auditability requires more than log files. It requires semantic traceability: the ability to reconstruct why a scenario was generated, which data were used, which agents dissented, and how conflicts were resolved. Principles alone cannot guarantee ethical AI because abstract commitments often fail to translate into operational controls [14]. Agentic systems therefore need concrete mechanisms such as role-based access, immutable records, challenge procedures, and independent validation. Human-centered AI research emphasizes reliable, safe, and trustworthy systems that keep humans in meaningful control [15]. In practice, this means designing interfaces that help risk officers understand agent reasoning, interrogate assumptions, and override recommendations without disabling the system's useful functions.

Accountability also has an institutional dimension. If multiple banks use similar agentic stress testing platforms, their scenarios and reactions may become correlated. This could create model monoculture, where institutions simultaneously reduce exposure to assets flagged by similar agents, amplifying fire sales and funding stress. Supervisors may need to monitor not only individual model risk but also the systemic effects of shared AI infrastructures. Governance should therefore include diversity requirements, independent scenario challenges, and supervisory visibility into common tools. The goal is not to suppress automation but to prevent automation from becoming a hidden coordination mechanism that undermines financial stability.

## **6. Robustness, Fairness, and Systemic Externalities**

Robustness in agentic stress testing means maintaining useful performance across data corruption, adversarial manipulation, regime change, and operational failure. Financial data are noisy, revised, and sometimes deliberately distorted. Agents that retrieve external text or market signals may be vulnerable to misinformation, spoofing, or prompt injection. They may also overfit to the most recent crisis or to the preferences of their operators. Robustness therefore requires adversarial testing, uncertainty quantification, redundant models, and graceful degradation. When an agent cannot establish reliable evidence, it should abstain, escalate, or produce a range of conditional outputs rather than a single authoritative answer.

Fairness enters stress testing in at least two ways. First, the distributional consequences of stress scenarios matter. A scenario that stabilizes aggregate capital may impose severe costs on particular regions, sectors, or households. Second, the use of AI in credit and risk decisions can produce discriminatory or exclusionary effects. Research on machine learning in credit markets shows that improved prediction can reshape access to finance and may have uneven distributional impacts [18]. Agentic systems could amplify these effects if they optimize for institutional loss minimization without considering broader social outcomes. Governance should therefore require distributional impact assessment, fairness testing, and stakeholder consultation where stress scenarios inform lending, forbearance, or resolution decisions.

Systemic externalities are especially important because stress testing is not a private forecasting exercise. It influences capital buffers, dividend policies, lending behavior, and market expectations. If agentic systems generate more frequent and granular stress signals,

they may encourage faster deleveraging during downturns, potentially increasing procyclicality. If they generate false alarms, they may impose unnecessary economic costs. If they miss slow-building vulnerabilities, they may create complacency. The design challenge is to balance timeliness with stability, sensitivity with specificity, and automation with discretion. Agentic AI should be evaluated by its contribution to the resilience of the financial system, not merely by the sophistication of its outputs.

## **7. Deployment, Infrastructure, and Sustainability**

Deployment of agentic stress testing is an infrastructure problem. It requires data platforms capable of serving high-frequency, heterogeneous inputs; compute environments for simulation and training; identity and access management; secure tool integration; and monitoring systems for performance and drift. Many financial institutions operate legacy systems, fragmented data warehouses, and strict change control processes. Introducing agentic AI into this environment can create integration risk, operational fragility, and hidden dependencies. The technical debt literature warns that machine learning systems accumulate entanglement and configuration debt that can undermine long-term reliability [11]. Agentic systems intensify this because they combine models, tools, memory, and autonomous decision loops.

Computational cost and energy use are also material. Large-scale scenario generation, multi-agent deliberation, and repeated simulation can be computationally intensive. If stress testing becomes continuous rather than periodic, the energy footprint may grow substantially. Sustainability considerations should therefore inform architecture choices, such as caching, approximate simulation, tiered escalation, and selective use of high-cost reasoning. Financial institutions and regulators may need to account for the environmental externalities of their risk infrastructure, especially if agentic systems are used for climate stress testing. A system that models climate risk but consumes excessive energy without justification presents a governance contradiction.

Deployment should also be staged. High-consequence use cases, such as setting capital buffers or triggering recovery plans, require stricter controls than exploratory research. Supervised sandboxes can allow institutions to test agentic capabilities without exposing live decision processes. Regulators may develop supervisory technology to inspect agent behavior, scenario provenance, and model dependencies. Cross-border coordination is necessary because financial stress does not respect jurisdictional boundaries. Infrastructure standards for data sharing, privacy, cybersecurity, and audit interoperability will shape whether agentic stress testing becomes a public good or a source of competitive fragmentation.

## **8. Policy and Institutional Implications**

Policy makers face a dual mandate: encourage innovation that improves financial resilience while preventing new sources of systemic risk. Agentic AI could enhance macroprudential surveillance by detecting emerging vulnerabilities, simulating cross-sector contagion, and stress testing policy responses. Supervisory stress tests have already evolved to include bank solvency, macro-financial feedback, and climate considerations [19]. Agentic systems could extend these efforts by generating adaptive scenarios and exploring nonlinear interactions. However, policy frameworks must address validation, transparency, and liability. If an agentic system is used in supervision, its outputs must be contestable by regulated institutions. If it is used internally, supervisors need sufficient access to assess reliability and comparability.

Goodhart's law warns that once a measure becomes a target, it ceases to be a good measure [20]. In stress testing, this risk appears when institutions optimize for passing a test rather than for genuine resilience. Agentic AI may accelerate such gaming if it learns to predict supervisory scenarios or to present portfolios favorably. Countermeasures include randomized scenario elements, independent validation, and penalties for superficial compliance. Policy should also encourage diversity in scenario generation and model architectures to reduce correlated behavior. A financial system in which all major institutions rely on the same few agentic platforms may be efficient in normal times but fragile under stress.

Institutional capacity is another constraint. Smaller banks, insurers, and public authorities may lack the expertise to build or supervise agentic systems. Without shared infrastructure or regulatory guidance, adoption could widen the gap between sophisticated and less sophisticated institutions. Public-private collaboration may be needed to develop common standards, reference scenarios, and audit tools. Universities, central banks, and international bodies can contribute through research, training, and open benchmarks. The policy objective should be to make agentic stress testing trustworthy, accessible, and accountable without turning it into a compliance ritual. That requires a governance culture that values challenge, uncertainty, and learning over automated certainty.

## **9. Conclusion**

Agentic AI has the potential to transform automated stress testing and financial risk scenario analysis by making scenario generation more adaptive, multi-perspectival, and responsive to regime shifts. It can help institutions explore tail events, climate transition pathways, cyber disruptions, and liquidity spirals that are difficult to capture with static models. Yet the same capabilities that make agentic systems powerful also create risks of opacity, correlated behavior, false confidence, and governance drift. The decisive issue is not whether agents can generate scenarios, but whether they can be embedded in architectures that preserve human responsibility, auditability, and systemic stability.

This paper has argued that agentic stress testing should be treated as critical socio-technical infrastructure. Its design must integrate data provenance, orchestration, trust calibration, adversarial validation, fairness assessment, and policy oversight. Autonomy should be graduated according to consequence, with clear escalation paths and contestability mechanisms. Institutions should avoid model monoculture, monitor systemic externalities, and account for the computational and environmental costs of continuous simulation. Regulators should develop supervisory capacity to inspect agent behavior and ensure that stress testing serves resilience rather than compliance theater. If governed carefully, agentic AI can deepen the financial system's ability to imagine and withstand future shocks. If governed poorly, it may become another source of fragility. The task is therefore as much institutional and political as it is technical.

## **References**

1. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
2. Wooldridge, M. (2009). *An Introduction to MultiAgent Systems* (2nd ed.). Wiley.
3. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
4. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe,

- D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
5. Bank for International Settlements. (2021). Annual Economic Report 2021. Bank for International Settlements.
  6. Financial Stability Board. (2020). The implications of climate change for financial stability. Financial Stability Board.
  7. Basel Committee on Banking Supervision. (2018). Stress testing principles. Bank for International Settlements.
  8. Liu, T. (2026). Trust-Calibrated Multi-Agent Decision Intelligence for Financial Risk Management under Market Regime Shifts. Available at SSRN 7307503.
  9. Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6), 1488-1542.
  10. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.
  11. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503-2511.
  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. 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. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44.
  14. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501-507.
  15. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe, and trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504.
  16. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291.
  17. Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.
  18. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5-47.
  19. European Central Bank. (2023). Macroprudential stress test of the euro area banking system. European Central Bank.
  20. Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. *Papers in Monetary Economics*, Reserve Bank of Australia.