Hedging the Black Box: A Feasibility Study of Financial Risk Transfer for AI Adoption via Generative Agent Simulation

Yixuan Yuan
Department of Data Science, University of Macau
Macau, China
yixuan.yuan@connect.um.edu.mo

Dedai Wei
Department of Economics, University of Macau
Macau, China
harrison.will@connect.um.edu.mo

Chudong Qian
Faculty of Business Administration, University of Macau
Macau, China
qianchudong@outlook.com

Ziyue Lin
School of Data Science, Fudan University
Shanghai, China
ziyuelin917@gmail.com

Yuheng Zhao
School of Data Science, Fudan University
Shanghai, China
yuhengzhao@fudan.edu.cn

Xinwu Ye*
School of Computing and Data Science, The University of Hong Kong
Hong Kong, China
xinwuye43@connect.hku.hk

ABSTRACT

The rapid evolution of artificial intelligence (AI) tools has demonstrated immense potential to enhance societal well-being and operational efficiency. However, the "black box" nature and inherent unreliability of modern AI systems, typified by large language models (LLMs), have created a significant barrier to entry. Many enterprises remain hesitant to integrate these tools deeply into their workflows due to concerns about unpredictable losses and liability. While existing technical solutions strive to enhance model reliability, they often struggle to effectively mitigate the "long-tail" risks associated with real-world deployment. In this paper, we introduce a socio-economic framework that allows enterprises to mitigate the stochastic risks of AI adoption by leveraging financial hedging instruments. To validate it, we introduce an LLM-driven Agent-Based Social Simulation (LABSS) system. We demonstrate the effectiveness of our simulation by benchmarking it against established economic and sociological baseline theories, as well as comparing it with traditional agent-based modeling (ABM). Our analysis demonstrates that the proposed financial framework effectively mitigates individual risk, thereby significantly accelerating the aggregate adoption rate of AI tools and promoting overall social productivity.

KEYWORDS

AI governance; LABSS; AI and Human Wellbeing

1 INTRODUCTION

AI has emerged as a transformative force in the modern economy, offering unprecedented opportunities to optimize operational workflows and elevate societal standards across diverse sectors [1, 9], while the integration of these systems into critical business infrastructure is hindered by their inherent unreliability [25, 27]. A significant dichotomy exists between the potential for aggregate social benefit and the localized, catastrophic risks faced by individual firms [24]. From a government macroeconomic perspective, AI adoption demonstrates a stable trajectory of increasing social benefit (Figure 1(a)). In contrast, individual enterprises are deterred by the looming threat of sudden, ruinous losses (Figure 1(b)) [14]. However, existing technical paradigms remain insufficient: as illustrated in Figure 1(c), "black box" scaling fails to eliminate long-tail risks, while "white box" approaches often sacrifice competitiveness by degrading model performance.

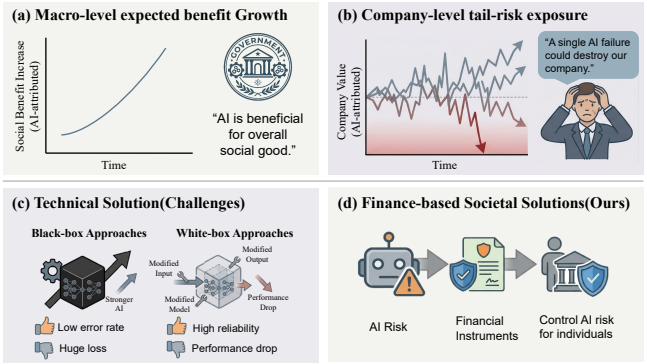

Figure 1: Motivation. (a)-(b) The misalignment between macro-level stability and micro-level volatility. (c)-(d) Comparison of risk mitigation paradigms.

Conventional narratives often assume that AI adoption is primarily a function of model capability [5], implying that accuracy improvements automatically drive integration. However, we propose that operational risk profiles, rather than model capabilities alone, often constitute the binding constraint on widespread commercial deployment. Therefore, we propose a solution rooted in a socio-economic perspective: utilizing financial instruments to hedge against the stochastic risks of AI failures (Figure 1(d)). This financial instrument functions by converting unbounded operational liabilities into manageable, fixed costs, as illustrated in Figure 2. By converting existential tail risks into manageable operational costs, our framework resolves the tension between micro-level risk aversion and the macro-level imperative for technological advancement [26].

To validate the efficacy of the proposed financial framework, we employ the **LLM-driven Agent-Based Social Simulation (LABSS)**. Unlike traditional ABM which is constrained by static rules, LABSS leverages generative cognitive architectures to capture the non-linear dynamics of risk perception and social contagion essential for realistic social simulation [10, 18, 19, 23]. Within this virtual society, heterogeneous firms interpret diverse environment information, such

*Corresponding author.

as vendor advertisements, accident reports, and insurance premium rates, to formulate strategic adoption and hedging commitments under uncertainty. This granular modeling captures the non-linear feedback loops between individual risk perception and system regimes. We evaluate our simulation against established sociological theories, such as the Fear Of Missing Out (FOMO) [6] and Prospect Theory [2], to ensure behavioral realism. Furthermore, we explicitly benchmark our approach against Rule-based ABM. This comparison proves that capturing cognitive dynamics is essential for reproducing realistic adoption trajectories, whereas deterministic logic fails to align with empirical observations.

As a framework demonstration, we conduct a comparative analysis against a control scenario lacking insurance mechanisms. Our analysis reveals that without financial hedging instruments, AI adoption rates stagnate as organizations remain paralyzed by potential liabilities. Conversely, the availability of these instruments functions as a critical enabler, significantly catalyzing technological diffusion by converting indefinite risks into manageable costs, thereby driving the enhancement of aggregate social productivity.

In summary, this paper makes the following key contributions:

- **Proposal of a Socio-Economic AI Risk Hedging Framework:** We introduce a interdisciplinary framework that utilizes financial instruments to hedge against stochastic algorithmic risks. This approach effectively resolves the tension between micro-level risk aversion and macro-level productivity to promote widespread AI adoption.
- **Development of the LABSS System:** We design and implement the LABSS. By incorporating generative agents with realistic cognitive architectures, LABSS captures the complex interactions between narrative propagation and decision-making.
- **Identification of Emergent Social Phenomena:** Through simulation, we uncover critical socio-economic dynamics emergent within this framework, offering nuanced insights for policymakers.

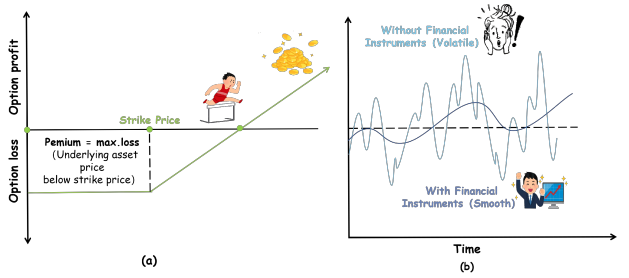

Figure 2: (a) Financial hedging principle, exemplified by American options. (b) Volatility-dampening effect of financial instruments.

2 RELATED WORK

LLM Agent-Based Social Simulation. The integration of LLMs shifts social simulation from rule-based to cognitive-based modeling. In micro-dynamics, studies demonstrate the emergence of scale-free networks [22] and norm propagation via observational learning [12], capturing non-linear behaviors. For phenomenon interpretation and theory validation, agents mimic adaptive epidemic behaviors [28], generate coherent long-term actions [23], and replicate the "Small-World Theory" [8]. Finally, in policy forecasting, frameworks like EconAgent [17] assess complex adaptive responses to interventions such as taxation.

AI Governance. AI governance balances risk and value across technical and societal dimensions. Technically, AI governance counters the "Black Box" opacity [7] by enforcing transparency and explainability [15, 16]. Societally, it orchestrates cooperation among heterogeneous actors (academia, government, industry) to ensure ethical alignment [11], balancing innovation risks with societal norms.

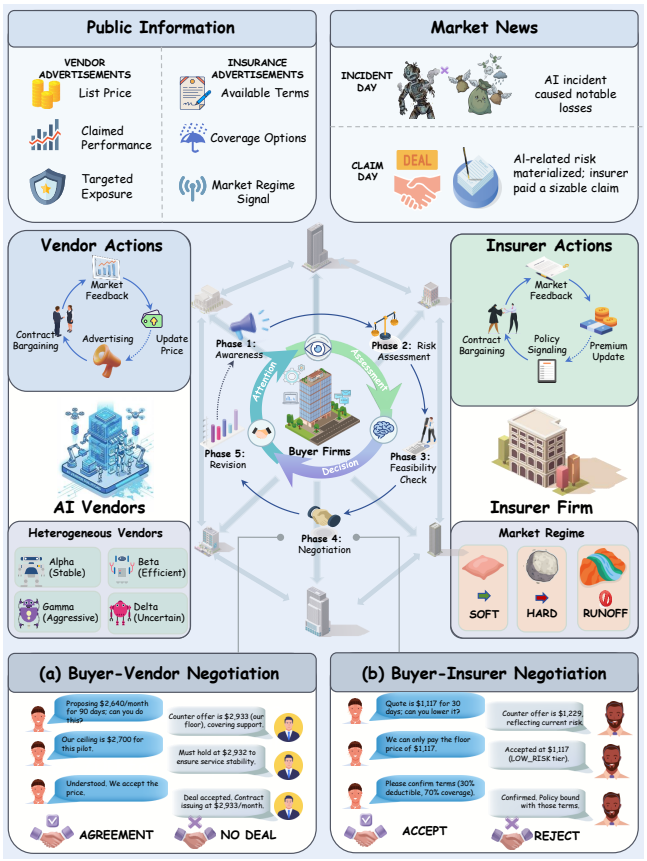

Figure 3: Overview. The system integrates macro-environmental uncertainty with micro-cognitive decision-making, orchestrating a lifecycle from signal processing to bilateral negotiation.

3 METHOD

We propose a risk hedging framework by introducing financial instruments to the society to mitigate AI adoption risks. To validate this framework, we employ insurance as the concrete realization of the financial hedging instrument, and construct a LABSS system. Specifically, we test the hypothesis that empowering firms to financially hedge against AI risk accelerates AI diffusion and enhances overall productivity in the simulation system. As illustrated in Figure 3, the framework integrates macro-environmental factors with micro-behavioral agents. This configuration ensures that the

aggregate diffusion of AI emerges directly from the continuous feedback loop between top-down institutional constraints and bottom-up entity choices.

3.1 Macro-Environmental Architecture

The macro-environmental architecture orchestrates the daily economic lifecycle through a structured interaction loop (Figure 4). The ecosystem is populated by three distinct categories of agents: buyer firms seeking productivity, AI vendors supplying technology, and a centralized insurer managing risk. This cyclical process governs how uncertainty is generated, information is disseminated, and contracts are ratified, unfolding sequentially across six operational phases.

Phase 1: Industry environment and uncertainty primitives. Buyer firms are stratified across heterogeneous industry sectors, mapped to different AI capabilities. On the supply side, vendor heterogeneity is characterized by distinct sector-specific specializations and variations in AI product performance. Based on this landscape, the insurer utilizes AI product profiles to establish dynamic premium rates that reflect the underlying AI risk.

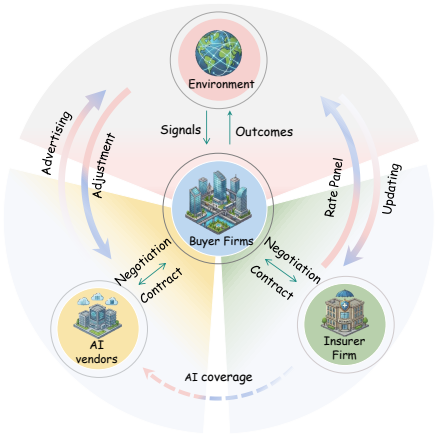

Figure 4: The Macro-Environmental Interaction Cycle. This diagram illustrates the cyclical flow of information and capital between the environment, AI vendors, insurers, and buyer firms, governing the transition from signal emission to contract realization.

Phase 2: Market publishing and information release. Daily interaction cycles begin with the dissemination of market signals. AI vendors advertise service profiles covering price, performance, and reputation, while the insurer updates coverage availability and sector-indexed quotes. Concurrently, to simulate real-world social dynamics, the environment propagates public and network-mediated narratives derived from recent realized events such as claims or exits, collectively defining the bounded information state available for firm observation.

Phase 3: Attention filtering and the endogenous visible set. Constrained by finite attention, the environment constructs buyer-specific consideration sets via a hybrid sampling mechanism that governs information access. This process allocates a uniform exploration slot to ensure non-zero discovery probabilities for niche vendors while filling remaining slots through weighted sampling proportional to vendor exposure. The resulting endogenous set defines the exclusive boundary for subsequent phase.

Phase 4: Action proposals under contractual constraints. Conditional on the visible set, firms formulate action plans regarding AI adoption and insurance hedging. These decisions are strictly gated by contract maturities where active terms enforce continuity while expiration windows authorize vendor switching or coverage adjustments. Validated intents are subsequently forwarded to the negotiation stage.

Phase 5: Negotiation and contract formation. Proposals are converted into binding commitments through negotiation. Firms bargain for service terms with AI vendors while insurance requests undergo underwriting subject to insurer constraints. Successful outcomes crystallize into formalized contracts specifying payment structures and coverage clauses whereas failed negotiations revert the firm to its status quo. These contracts define the governing rules for the subsequent realization of financial outcomes and claim assessments.

Phase 6: Settlement, outcome realization, and claims. The cycle concludes with financial settlement where firms remit fees and premiums. Operational performance is modeled as a composite process. It begins with a baseline return, $r_{i,t}^{\text{base}}$, which captures standard industry trends and idiosyncratic shocks independent of AI adoption. If a firm integrates AI (denoted by the indicator $\mathbf{1}\{\text{AI}_{i,t} = 1\}$), its performance is further modulated by AI performance, which is mapped into operational impacts, generating AI gains, $\text{gain}_{i,t}$, or defining convex penalties, $\text{loss}_{i,t}$. The final realized return for firm i integrates these components:

$$r_{i,t} = r_{i,t}^{\text{base}} + \mathbf{1}\{\text{AI}_{i,t} = 1\}(\text{gain}_{i,t} - \text{loss}_{i,t}) + \varepsilon_{i,t}. \tag{1}$$

Regarding risk transfer, verified claims result in indemnity payouts subject to policy limits and insurer solvency, effectively updating the market information state for the subsequent period.

3.2 Micro-Behavioral Dynamics

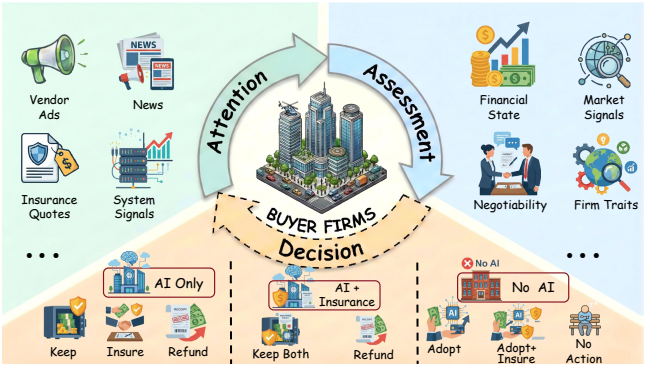

Figure 5: The AAD Framework. The micro-level cognitive process where firms filter signals (Attention), evaluate risks (Assessment), and execute strategies (Decision).

At the micro level, we model firm behavior through the lens of corporate behavioral finance and organizational theory. As illustrated in Figure 5, heterogeneous firms operate as boundedly rational entities utilizing the Attention-Assessment-Decision (AAD) framework.

Attention. Attention governs the availability of market information. Rather than assuming full omniscience, firms observe a limited and time-varying consideration set comprising visible AI advertisements, insurance quotes, and salient network signals. This exposure is strictly bounded by finite cognitive capacity, ensuring that firms process only a fraction of available opportunities. Consistent with endogenous consideration models, this mechanism acts as the primary filter determining which options enter subsequent evaluation [13].

Assessment. Assessment translates the filtered attention set into perceived risk and strategic intent. Driven by organizational risk pressure rather than deterministic profit maximization, this mechanism synthesizes internal factors like capital solvency and behavioral inertia with external network signals regarding market volatility. Consistent with corporate behavioral finance, the process yields discrete action tendencies ranging from status quo maintenance to AI termination or risk transfer.

Decision. Decision translates the preferences based on the assessment into binding operational commitments. The outcome determines the firm's specific operational mode: operating without AI, utilizing AI independently, or combining AI adoption with insurance hedging. The execution of these choices is conditional on negotiation success and market availability, effectively updating the firm's technological and contractual state for subsequent periods.

4 EXPERIMENTAL VALIDATION

We validate the simulation framework by verifying its alignment with established sociological theories at both micro and macro levels. Concurrently, we benchmark the LABSS system against rule-based ABM, concluding with sensitivity analyses to confirm the structural robustness of our findings.

4.1 Experimental Setup

To validate the proposed framework, we instantiate the LABSS environment as follows.

Simulation Environment. The simulation is initialized with 11 industry sectors aligned with the GICS standard [20]. The time granularity is set to daily interaction cycles spanning a total of 100 days.

AI Performance Simulation. Acknowledging that AI risk pricing is pivotal to the practical realization of financial instruments, we first simulate the prediction performance of AI models using real-world industry data as an AI risk proxy. Based on these empirical risk profiles, we propose a proof-of-concept pricing method grounded in the Diffusion Model and the Cox-Ross-Rubinstein Binomial Tree, which is used as the reference pricing for insurer in our simulation.

Agent Configuration. The economy consists of 300 firm agents driven by DeepSeek-V3.2. Each agent is initialized with unique, heterogeneous profiles derived from their industry characteristics, such as risk preferences and organizational inertia.

4.2 Simulation Effectiveness Validation

To validate the behavioral realism of the LABSS system, we compare simulation outcomes with established economic and sociological theory baselines, including the Bass Diffusion Model [4], FOMO [6], Path Dependence [2], and Prospect Theory [2].

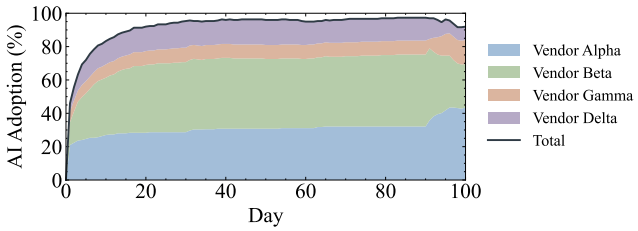

Figure 6: AI Adoption by Vendor.

Bass Diffusion Model and Path Dependence. We observe that AI penetration exhibits an explosive initialization (Figure 6), consistent with the high-innovation regime of the **Bass Diffusion Model**. As further validation, through micro-level analysis of agent cognitive logs, we identify distinct thinking patterns that underpin this model. Specifically, early adopters rationalize decisions as seizing a "competitive edge" (the "Innovator" profile), whereas later entrants override conservatism due to "social exclusion" fears, confirming the transition to **Herd Behavior**. Finally, the market stabilizes into a fixed distribution governed by **Path Dependence**, where contractual constraints create a feedback loop of Cognitive Lock-in [21].

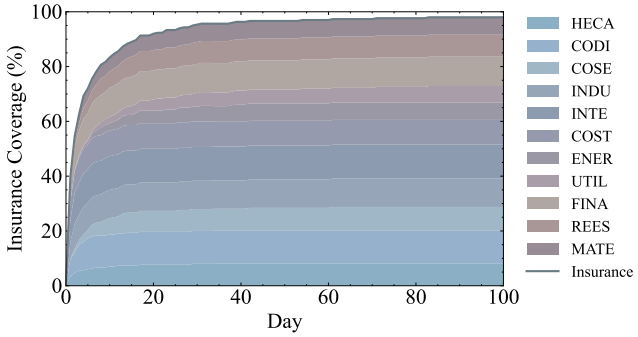

Figure 7: Insurance Coverage by Industry.

Prospect Theory. Insurance coverage demonstrates a rapid and comprehensive saturation across all sectors (Figure 7). This universal demand for stability directly validates the **Certainty Effect** within **Prospect Theory**. Through case analysis, we identify a representative behavioral pattern where risk-averse firms pay a "safety premium" for psychological relief rather than actuarial necessity. This finding confirms that our system effectively simulates non-rational behavioral anomalies, demonstrating how agents prioritize the elimination of catastrophic uncertainty over strict profit maximization.

4.3 ABM Baseline Comparison

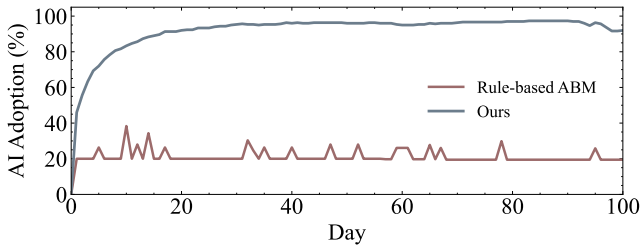

Figure 8: Comparison of AI Adoption Rates.

To validate the fidelity of our simulation, we benchmark our LLM-driven framework against a Rule-based ABM baseline governed by absolute rationality. The divergence in adoption trajectories, presented in Figure 8, serves as a critical validity test. The baseline fails to replicate the empirically observed technology diffusion, exhibiting instead a mechanistic collapse to a low equilibrium. This implies that agents driven solely by strict utility maximization prematurely exit the market upon detecting initial risks, contradicting the real-world proliferation of AI. This demonstrates that the inclusion of social contagion and behavioral resilience is essential for capturing the true dynamics of AI adoption.

4.4 Sensitivity Analysis

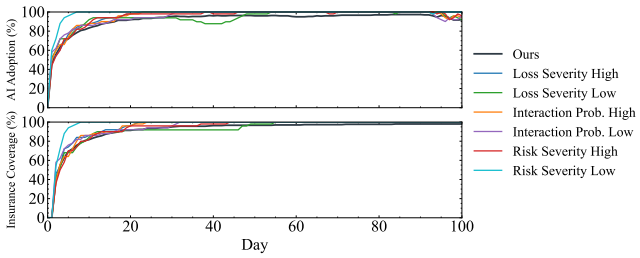

Figure 9: Robustness Analysis.

The structural stability of the proposed framework is corroborated by the sensitivity analysis in Figure 9, where the system converges to a high-saturation equilibrium across a wide range of parameter configurations. Perturbing critical variables ranging from the speed of social contagion to the magnitude of financial shocks results in only marginal deviations from the baseline trajectory. This consistent resilience suggests that the "Solvency Filter" effect of insurance is a fundamental structural property of the ecosystem rather than an artifact of specific hyperparameter tuning, validating that financial hedging remains an effective catalyst for AI diffusion even under varying conditions of information scarcity or heightened environmental volatility.

5 CAN THE INTRODUCTION OF FINANCIAL INSTRUMENTS PROMOTE AI ADOPTION?

In this section, we evaluate the causal impact of risk transfer mechanisms. We simulate under two opposing conditions: with and without the availability of financial hedging instruments, and analyze the resulting difference.

5.1 Risk Hedging Overcomes the "Adoption Ceiling"

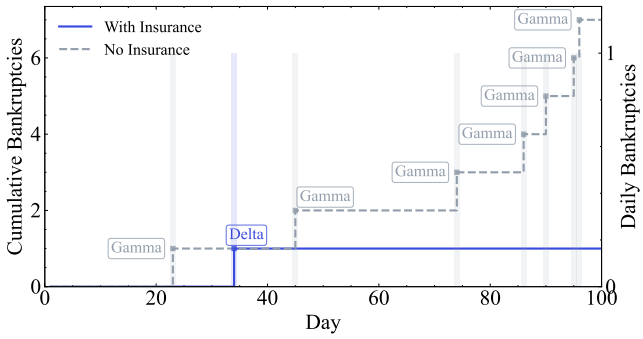

Figure 10: Survival Resilience and Solvency Filtering.

The most immediate impact of financial coverage is the enhancement of systemic survivability, as illustrated in Figure 10. In the No-Insurance scenario, the system operates as a survival of the fittest environment where idiosyncratic shocks lead directly to liquidity crises, resulting in 7 bankruptcies at Day 100. In contrast, the insured scenario limits bankruptcies to only 1 firm because the insurance mechanism acts as a critical solvency filter. This mechanism decouples the operational utility of AI from its inherent statistical risks by converting fatal tail events that would otherwise deplete capital buffers into manageable operational expenses like premiums and deductibles, thus preventing premature market exit due to stochastic shocks rather than strategic failures.

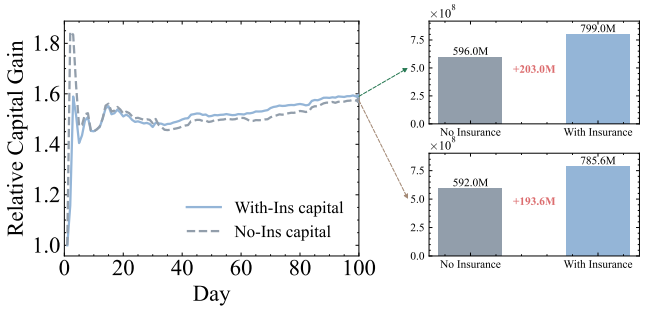

Figure 11: Net Welfare Effect Analysis.

To quantify the net welfare effect, we analyze the relative capital gain trajectory (Figure 11). Crucially, both the total societal wealth and the real economy (excluding insurer profits) consistently exceed 1.0, indicating that the insured ecosystem outperforms the baseline. Validated by the Day 100 snapshot where Total Social Capital significantly expands (7.81M vs. 5.96M), this confirms the existence of a "Stability Dividend." Financial hedging functions as a "Solvency Filter" that prevents idiosyncratic shocks from escalating into capital-destroying bankruptcies, allowing firms to survive and compound earnings beyond the direct costs of risk transfer. The in-

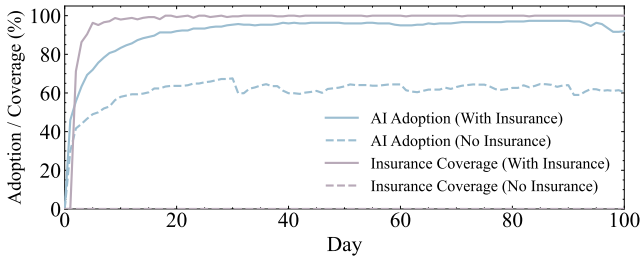

Figure 12: Comparison of AI Adoption Trajectories.

troduction of financial hedging instruments fundamentally alters the dynamics of technological diffusion by removing the structural barriers to entry for risk-averse agents (Figure 12). In the No Insurance scenario, market penetration stagnates at a distinct "adoption ceiling" (approximately 60%), a plateau driven by the rational paralysis of firms facing unbounded "black box" liabilities. Conversely, the With Insurance ecosystem exhibits a rapid convergence to near-universal saturation (> 95%). This stark divergence indicates that the primary impediment to widespread AI deployment is not a deficiency in operational utility, but the presence of unmanaged stochastic risks. By converting prohibitive tail exposures into deterministic operating costs (premiums), insurance effectively unlocks latent demand, allowing even conservative enterprises to integrate AI into core operations without the threat of existential ruin.

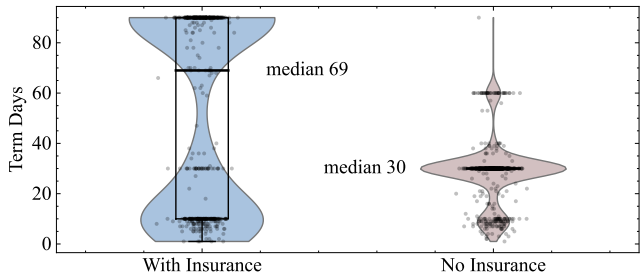

Figure 13: Distribution of contractual term lengths comparing the With Insurance and No Insurance scenarios.

Risk transfer mechanisms significantly alter the temporal structure of commercial commitments (Figure 13). Without insurance, firms exhibit a defensive operational posture, heavily favoring short-term engagements (median 30 days) to retain the option of rapid exit in response to potential algorithmic failures. This "testing the waters" approach reflects a lack of confidence in long-term stability. In contrast, the availability of coverage encourages a shift towards longer strategic horizons, with the median contract duration extending more than double to 69 days. This elongation suggests that when potential liabilities are capped, firms are empowered to prioritize operational continuity over short-term risk avoidance, fostering a more stable commercial environment conducive to deep technological integration.

5.2 Network Effects as the Diffusion Engine

Social contagion acts as the primary accelerator for market saturation, creating a powerful herding effect that amplifies the utility of insurance.To isolate the role of social contagion, the ablation study in Figure 14 severs the communication channels between agents. In

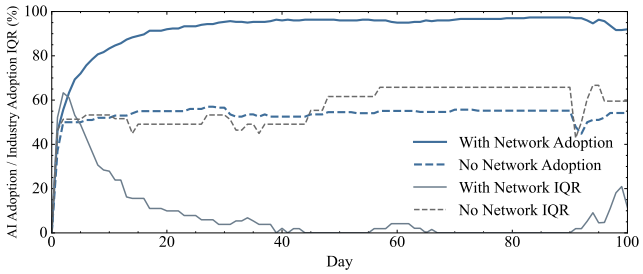

Figure 14: Ablation Study on Network Effects.

the With Network environment, the rapid saturation of AI adoption coincides with a collapsing Interquartile Range (IQR), indicating that information diffusion creates a powerful herding effect where diverse industries converge onto a unified strategy in lockstep. Conversely, the removal of network signals results in a sluggish adoption trajectory accompanied by a persistently high IQR. This increased dispersion reveals that without the unifying pressure of social proof, decision-making reverts to idiosyncratic internal constraints, resulting in a fragmented market where sectors diverge significantly rather than coalescing into a systemic consensus.

5.3 Allocative Inefficiency

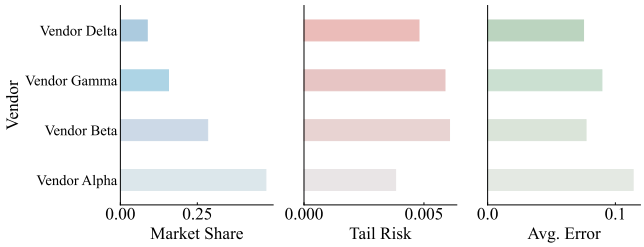

Figure 15: The Gold Plating Inefficiency.

While financial hedging effectively mitigates AI deployment risks, it simultaneously induces a "Gold Plating" inefficiency [3], where stability is prioritized over predictive productivity (Figure 15).Vendor Alpha achieves a dominant market share despite not being the most accurate model available; its dominance stems solely from having the lowest tail risk profile rather than superior predictive performance. In a perfectly efficient market, firms would likely converge on high-accuracy competitors like Vendor Delta while using insurance to manage tail risks. However, the simulation reveals a strong bias for absolute safety, where the availability of insurance paradoxically reinforces a conservative clustering around the safest incumbent. The net result is allocative inefficiency, as the economy overspends on safety premiums while stifling the adoption of more efficient technologies.

The decomposition of aggregate market saturation in Figure 16 reveals how distinct risk profiles dictate vendor selection. High-stakes sectors exhibit a pronounced concentration around the stability-focused Vendor Alpha, validating the preference for minimized tail risk over raw predictive performance. Crucially, the sparse distribution of uninsured firms, represented by the dark grey segments,

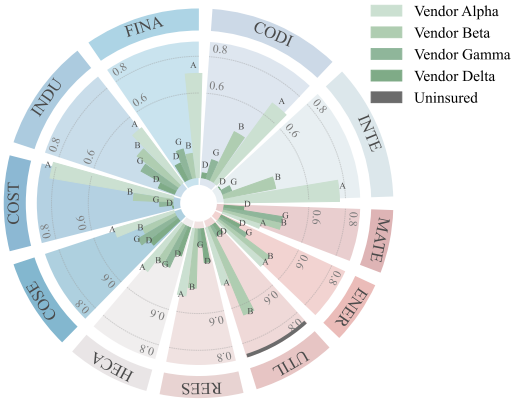

Figure 16: Sector-Specific Vendor Selection and Insurance Penetration.

underscores that risk transfer is not merely an option but a structural necessity across the board. This omnipresence of coverage, regardless of the specific vendor chosen, confirms that financial hedging serves as the fundamental "license to operate" that unifies an otherwise fragmented landscape of operational requirements.

5.4 The Decoupling of Risk Perception from Reality

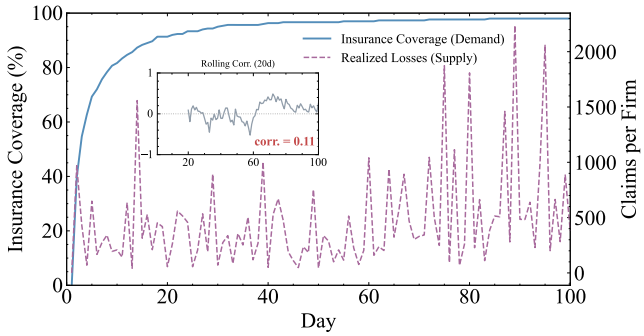

Figure 17: Comparison between the aggregate insurance demand and the fundamental supply of risk events. The inset chart tracks the rolling correlation coefficient between them.

The structural decoupling of risk perception from material reality is quantitatively evidenced by a negligible correlation coefficient ($\rho = 0.11$) between the aggregate demand for insurance and actual realized losses (Figure 17). This statistical disconnect signals a side effect of financialization where insurance functions less as a rational hedge against tangible operational risks and more as an "emotional placebo" or institutional prerequisite. The macroeconomic consequence is a misallocation of social capital where a certain amount of financial resources are siphoned from productive activities into unnecessary premiums. Instead of fueling innovation, this capital stagnates as idle solvency reserves within the insurance sector effectively creating a drag on the real economy.

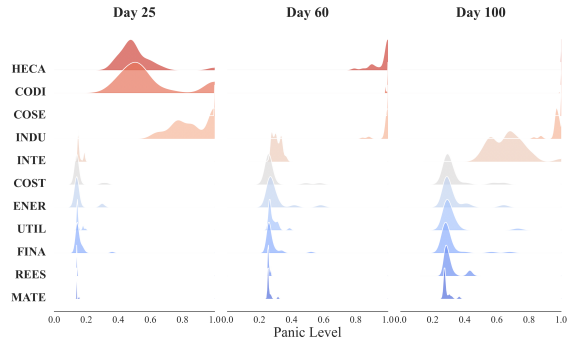

Figure 18: Structural Evolution of Panic Distribution.

5.5 The Structural Evolution of Systemic Panic

The temporal dynamics of risk sentiment reveal a transition from uniformity to complexity, where Figure 18 tracks the structural differentiation that emerges endogenously across industries. Initially at Day 25, the system exhibits a high degree of homogeneity with distributions concentrated near zero. However, as the simulation progresses, vulnerable industries such as Communication Services and Financials develop pronounced heavy tails and bimodal distributions, signaling a polarization where a subset of firms enters a high-stress state while others remain resilient. In contrast, robust sectors like Utilities maintain a flat and low-risk profile. Ultimately, this divergence signifies the crystallization of a stratified risk landscape, where inherent sector-specific vulnerabilities dictate the distinct evolutionary paths ranging from stability to extreme fragility, rather than a uniform market-wide diffusion.

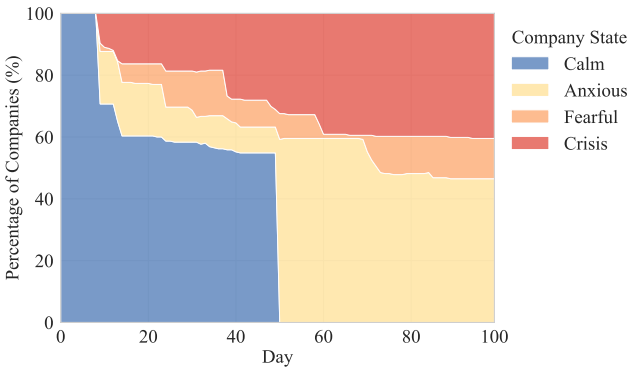

Figure 19: Evolution of Buyer Firm Sentiment Regimes.

The aggregate sentiment of buyer firms demonstrates a clear nonlinear accumulation of systemic risk (Figure 19). The initially dominant Calm state is progressively eroded by the expansion of Fearful and Crisis regimes. Crucially, the ecosystem does not revert to its initial equilibrium following shock events. Instead, it settles into a "New Normal" characterized by a persistent structural component of distressed firms, comprising approximately 20% to 30% of the population. This hysteresis effect reveals that the interaction between algorithmic failures and liability mechanisms creates a feedback

loop where risk is redistributed rather than fully absorbed, leading to a lasting transformation in the macro-state of the ecosystem.

6 CONCLUSION

This study validates that the primary barrier to widespread AI utility is not only model capacity but the financial capacity of enterprises to absorb tail risk. By introducing a financial hedging framework, we demonstrated that converting stochastic "black box" volatility into manageable premiums acts as a solvency filter, successfully driving AI adoption from a stagnant 60% to near-universal saturation in our LABSS. However, the simulation also uncovers emergent pathologies, specifically a "Gold Plating" inefficiency where firms, driven by intrinsic loss aversion, prioritize stability over innovation, leading to a structural decoupling between capital allocation and productive optimization. These findings confirm the efficacy of financial instruments in bridging the gap between risk-averse enterprises and technological advancement while highlighting the critical need for future governance mechanisms to balance systemic stability with allocative efficiency.

REFERENCES

[1] Daron Acemoglu. 2025. The simple macroeconomics of AI. *Economic Policy* 40, 121 (2025), 13–58.

[2] W Brian Arthur. 1989. Competing technologies, increasing returns, and lock-in by historical events. *The economic journal* 99, 394 (1989), 116–131.

[3] Harvey Averch and Leland L Johnson. 1962. Behavior of the firm under regulatory constraint. *The American Economic Review* 52, 5 (1962), 1052–1069.

[4] Frank M Bass. 1969. A new product growth for model consumer durables. *Management science* 15, 5 (1969), 215–227.

[5] Patrick Bedué and Albrecht Fritzsche. 2022. Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management* 35, 2 (2022), 530–549.

[6] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* 100, 5 (1992), 992–1026.

[7] Alexander Buhmann and Christian Fieseler. 2021. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society* 64 (2021), 101475.

[8] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*. 3326–3346.

[9] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130* 10 (2023).

[10] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.

[11] Urs Gasser and Virgilio AF Almeida. 2017. A layered model for AI governance. *IEEE Internet Computing* 21, 6 (2017), 58–62.

[12] Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh. 2024. Generative agent-based modeling: an introduction and tutorial. *System Dynamics Review* 40, 1 (2024), e1761.

[13] John R Hauser and Birger Wernerfelt. 1990. An evaluation cost model of consideration sets. *Journal of consumer research* 16, 4 (1990), 393–408.

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001* (2023).

[15] Joshua A Kroll. 2018. The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180084.

[16] Stefan Larsson. 2020. On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society* 7, 3 (2020), 437–451.

[17] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. 15523–15536.

[18] Ziyue Lin, Yi Shan, Lin Gao, Xinghua Jia, and Siming Chen. 2025. SimSpark: Interactive Simulation of Social Media Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–32.

[19] Ziyue Lin, Siqi Shen, Zichen Cheng, Cheok Lam Lai, and Siming Chen. 2025. Carbon and silicon, coexist or compete? a survey on human-ai interactions in agent-based modeling and simulation. *arXiv preprint arXiv:2502.18145* (2025).

[20] SP MSCI. 2006. The Global Industry Classification Standard (GICS®).

[21] Kyle B Murray and Gerald Häubl. 2007. Explaining cognitive lock-in: The role of skill-based habits of use in consumer choice. *Journal of Consumer Research* 34, 1 (2007), 77–88.

[22] Marios Papachristou and Yuan Yuan. 2025. Network formation and dynamics among multi-llms. *PNAS nexus* 4, 12 (2025), pgaf317.

[23] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[24] Dragutin Petkovic. 2023. It is not "Accuracy vs. Explainability"—we need both for trustworthy AI systems. *IEEE Transactions on Technology and Society* 4, 1 (2023), 46–53.

[25] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[26] Philip Moreira Tomei, Rupal Jain, and Matija Franklin. 2025. AI governance through markets. *arXiv preprint arXiv:2501.17755* (2025).

[27] Warren J Von Eschenbach. 2021. Transparency and the black box problem: Why we do not trust AI. *Philosophy & technology* 34, 4 (2021), 1607–1622.

[28] Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986* (2023).