International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 3, 169-190, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P118

Original Article

The Trust Threshold: How Public Perception of AI Harm Moderates the Impact of FinTech Innovation on Systemic Banking Stability

Rajitha Gentyala
Frisco, Texas, USA.

Abstract - *The rapid integration of artificial intelligence into financial services has ushered in an era of unprecedented innovation, yet it has simultaneously introduced complex socio-technical risks that challenge conventional understandings of banking stability. While substantial research has examined the technical dimensions of AI model governance and regulatory compliance, comparatively little attention has been devoted to understanding how public perceptions of AI-related harm influence the relationship between technological innovation and systemic financial resilience. This study addresses this critical gap by investigating the moderating role of public trust and perceived socio-political harm in the innovation-stability nexus within the banking sector. Drawing upon diffusion of innovation theory and systemic risk frameworks, we develop and test a conceptual model wherein public perception of AI harm—encompassing concerns regarding algorithmic fairness, data privacy, and opaque decision-making—moderates the impact of aggressive AI adoption on long-term banking stability. The research employs a mixed-methods approach, combining longitudinal analysis of banking stability indicators across major financial institutions with survey data capturing public sentiment toward AI deployment in financial services. Preliminary findings suggest that while AI innovation initially enhances operational efficiency and profitability, these benefits are contingent upon maintaining public confidence in the fairness and integrity of automated systems. When perceptions of harm exceed a critical threshold, the stability benefits of innovation diminish significantly, potentially triggering customer attrition, regulatory intervention, and systemic contagion effects. The study draws upon foundational insights from Cao et al. (2021), who examined consumer trust dynamics in algorithmic financial advising, and extends the work of König et al. (2022), who explored the reputational contagion mechanisms linking perceived AI failures to broader institutional stability. By illuminating the psychological and sociological dimensions of AI governance, this research contributes to emerging scholarships on trustworthy AI and offers practical guidance for financial institutions seeking to balance innovation imperatives with the maintenance of public trust. The findings underscore the necessity of embedding public perception monitoring into systemic risk assessment frameworks and highlight the importance of transparent, explainable AI architectures in preserving the social license upon which banking stability ultimately depends.*

Keywords - *Artificial Intelligence, Banking Stability, Public Perception, Algorithmic Harm, Financial Innovation, Systemic Risk, Trust Threshold, Socio-Political Risk, AI Governance, Consumer Confidence*

1. Introduction

The integration of artificial intelligence into financial services represents one of the most significant transformations in modern banking history, reshaping everything from customer interactions and credit decisions to risk management and regulatory compliance. As financial institutions increasingly delegate critical functions to algorithmic systems, questions once reserved for human judgment are now being answered by machine learning models operating at scales and speeds impossible for human practitioners to match. This technological evolution has generated substantial efficiency gains, enabling faster loan approvals, more sophisticated fraud detection, and personalized financial advice delivered through robo-advisors and chatbots. Yet beneath these operational improvements lies a more complex and less understood dynamic: the relationship between how the public perceives AI-driven financial services and the long-term stability of the banking systems that increasingly depend upon them. The present study addresses this gap by investigating the moderating role of public perception of AI harm in the relationship between financial innovation and systemic banking stability, proposing that trust thresholds may fundamentally condition whether technological advancement strengthens or undermines the resilience of financial institutions.

The scholarly literature examining AI in banking has developed along two relatively distinct trajectories that have seldom intersected in meaningful ways. On one hand, substantial research has focused on the technical dimensions of AI governance, including model validation, data quality assurance, explainability requirements, and compliance with emerging regulatory frameworks. This body of work has produced important insights into how financial institutions can develop and deploy AI systems that meet technical standards for accuracy, fairness, and transparency. On the other hand, a separate stream of research has examined consumer adoption of financial technologies, identifying factors that influence whether individuals choose to use robo-advisors, mobile banking applications, and algorithmically mediated financial services. What remains comparatively underdeveloped is systematic investigation of how public perceptions of AI-related harm might moderate the relationship between innovation and stability, potentially serving as a mechanism through which individual attitudes aggregate to system-level consequences.

Aitken et al. [1] provides foundational insights into this underexplored territory through their qualitative study of public attitudes toward AI in banking. Conducting focus groups with diverse members of the public, these researchers uncovered a phenomenon they describe as cognitive dissonance, wherein participants readily used AI-powered financial services for reasons of convenience and immediate benefit while simultaneously expressing profound distrust and concern about the societal implications of these same technologies. Notably, the concerns articulated by participants did not typically center on private or individual interests but rather extended to wider ethical and social considerations, including anxiety about algorithmic bias, the erosion of human judgment in consequential decisions, and the concentration of power in increasingly automated financial systems. The study's emphasis on conditions for public acceptability rather than mere customer uptake suggests that banking institutions cannot assume that usage patterns reflect genuine trust or sustainable confidence in AI-driven services. This distinction carries important implications for stability, as the cognitive dissonance identified by Aitken et al. [1] may render the relationship between innovation and public confidence brittle, susceptible to rapid deterioration when perceived harm materializes or becomes salient.

The regulatory landscape surrounding AI in financial services has evolved considerably in response to these concerns, though significant gaps remain in how systemic risks are conceptualized and addressed. PwC [2] documents the accelerating pace of AI adoption across the financial sector, noting that a 2023 survey by the Bank of England and Financial Conduct Authority revealed 72 percent of firms are using or developing machine learning applications, with firms anticipating a 3.5-fold increase in such applications by 2026. This dramatic growth has attracted increasing regulatory scrutiny, with authorities in the United Kingdom, European Union, and other jurisdictions developing frameworks intended to ensure responsible AI deployment. However, as PwC [2] observes, the regulatory conversation has focused predominantly on consumer protection, operational resilience, and model risk management at the institutional level, with comparatively less attention devoted to the systemic implications of widespread AI adoption or the potential for public perception to function as a transmission mechanism for financial contagion. The emerging debate over risks associated with generative AI and artificial general intelligence has begun to encompass systemic and longer-term concerns related to market resilience and large-scale misinformation, yet these discussions remain in early stages and have not yet been fully integrated into stability monitoring frameworks.

The theoretical motivation for this study emerges from the intersection of these literatures and from the recognition that banking stability is not merely a technical condition measurable through capital ratios and liquidity metrics but is fundamentally shaped by psychological and social dynamics. Classic bank run theory demonstrates that confidence functions as a critical stability determinant, with losses of trust capable of triggering self-fulfilling prophesies that transform individual withdrawals into systemic crises. In the context of AI-driven banking, this insight acquires new urgency because the mechanisms through which trust is established and maintained differ substantially from those operating in traditional human-mediated financial relationships. When customers interact with algorithmic systems, they cannot draw upon the same social cues, relational histories, or expectations of accountability that characterize interactions with human bankers. Perceptions of fairness, interpretability, and trustworthiness become filtered through assumptions about automation that may systematically disadvantage algorithmic systems relative to human alternatives, even when objective performance is equivalent or superior.

The research gap addressed by this study is therefore both timely and consequential. While existing scholarship examined either AI governance or banking stability independently, and while valuable work has explored individual adoption decisions, the field lacks systematic investigation of how public perception of AI harm may moderate the aggregate relationship between innovation and stability. Understanding this moderation effect requires theoretical frameworks that can bridge micro-level psychological processes and macro-level systemic outcomes, as well as empirical strategies capable of capturing complex dynamics through which individual perceptions scale to system-level consequences. The present research responds to this gap by proposing that trust thresholds exist, points at which the accumulation of perceived harm or the occurrence of salient failures triggers shifts in public confidence with measurable implications for banking stability. By integrating insights from Aitken et al. [1] on the cognitive dissonance characterizing public attitudes and from PwC [2] on the regulatory and adoption landscape, this study

develops a conceptual framework for understanding how the social license under which AI operates in banking may prove more fragile than current governance approaches assume.

The structure of this literature review proceeds as follows. Section II examines theoretical foundations of trust in automated financial systems, drawing upon psychological and sociological frameworks to explain how humans develop confidence in algorithmic decision-making. Section III systematically categorizes the multidimensional nature of perceived AI harm in banking, distinguishing between fairness concerns, privacy anxieties, opacity problems, and systemic risks. Section IV reviews empirical evidence on how AI innovation has affected banking performance, establishing the baseline relationship that public perception may subsequently moderate. Section V develops the conceptual framework linking public perception to systemic stability through customer behavior, regulatory response, and contagion dynamics. Section VI synthesizes these literatures to identify specific gaps and justify the current study's contribution to understanding how trust thresholds condition the relationship between AI innovation and banking stability.

2. Theoretical Foundations of Trust in Automated Financial Systems

Understanding how trust operates in the context of automated financial services requires a multidisciplinary theoretical lens that integrates insights from psychology, human-computer interaction, and technology acceptance research. Unlike traditional banking relationships, wherein trust develops through interpersonal interactions, repeated face-to-face encounters, and the cultivation of personal rapport with human bankers, AI-mediated financial services present a fundamentally different trust calculus. Customers interacting with robo-advisors, chatbots, and algorithmic lending platforms cannot observe facial expressions, interpret tone of voice, or draw upon shared social histories. Instead, they must form judgments about the competence, reliability, and goodwill of systems whose inner workings remain largely opaque and whose decision-making processes resist straightforward human interpretation. This section examines the theoretical foundations that explain how trust develops in such contexts, drawing upon contemporary research that has substantially advanced scholarly understanding of this phenomenon.

The conceptualization of trust in automated systems has evolved considerably from early unidimensional models toward more nuanced frameworks that recognize the multidimensional nature of trusting beliefs. Traditional trust theory, originating in interpersonal psychology, identified three foundational dimensions that collectively determine whether one party places confidence in another: ability, referring to the perceived competence and expertise of the trustee; integrity, encompassing the belief that the trustee adheres to principles acceptable to the trustor; and benevolence, reflecting the perception that the trustee cares about the trustor's welfare and acts in their interest rather than solely pursuing self-interest. These dimensions have proven remarkably durable across diverse contexts, yet their applicability to human-machine relationships cannot be assumed without empirical validation. The absence of intentionality in algorithmic systems raises fundamental questions about whether concepts like benevolence, which imply motivational states and conscious concern for others, can meaningfully apply to entities that lack consciousness or moral agency.

Schütz, Schröder, and Rennhak [3] provide critical empirical evidence addressing precisely this question through their experimental investigation of trust attributes in robo-advisor acceptance. Recognizing that automated investment advisory services lack human interaction traditionally considered essential for trust development, these researchers designed a study to examine how the classic trust dimensions of ability, integrity, and benevolence influence whether individuals follow investment recommendations generated by algorithmic systems. Their findings reveal a nuanced pattern with significant theoretical implications. Both ability and integrity emerged as significant predictors of recommendation adherence, indicating that consumers evaluate robo-advisors based on perceived competence and the expectation that these systems operate according to reliable, consistent principles. However, benevolence did not achieve statistical significance in their model, suggesting that this dimension may function differently when the trustee is an algorithm rather than a human advisor [3].

This differential functioning of benevolence warrants careful theoretical interpretation. Schütz et al. [3] propose that consumers may not expect or attribute benevolent motivations to algorithmic systems in the same way they would to human advisors. When interacting with a human financial advisor, clients reasonably wonder whether recommendations serve the client's interests or merely maximize the advisor's commission. This question of motive is fundamental to interpersonal trust. With algorithmic systems, however, consumers may recognize that machines lack intentions altogether, rendering the question of benevolent motivation conceptually incoherent. What matters instead is whether the algorithm has been designed and programmed to optimize outcomes aligned with user interests, a characteristic that arguably relates more closely to integrity and ability than to benevolence as traditionally conceived. This insight carries important implications for both theory and practice, suggesting that trust in AI may rest on a narrower foundation than interpersonal trust, one that emphasizes technical competence and reliable performance over perceived goodwill.

The theoretical picture becomes considerably more complex when considering research that has found significant effects for benevolence in AI contexts, suggesting that contextual and methodological factors may moderate which trust dimensions prove salient. Chang, Park, and Dinh [4] offer a compelling counterpoint through their application of Social Cognitive Theory to understand trust formation in AI-enabled financial services. Drawing upon Bandura's foundational framework, these researchers argue that trust serves as a precondition for vicarious learning, the process through which individuals observe and internalize the actions and outcomes of others to guide their own behavior. In the context of robo-advisors, vicarious learning occurs when consumers observe the advisory process, interpret the recommendations offered, and gradually develop confidence in their own ability to navigate financial decisions with algorithmic assistance. This theoretical framing positions trust not merely as an attitude toward technology but as an active cognitive mechanism that enables skill development and self-efficacy enhancement.

Chang et al. [4] distinguish between two trust dimensions that map partially but not perfectly onto the classic triad: credibility-based trust, encompassing beliefs about the AI advisor's competence and reliability, and benevolence-based trust, reflecting perceptions that the system acts with the consumer's best interests in mind. Their structural equation modeling analysis, conducted with data from 361 United States consumers, reveals that both trust dimensions positively influence financial self-efficacy, which in turn drives adoption intentions. Notably, benevolence exerted significant effects in their model, apparently contradicting the null findings reported by Schütz et al. [3]. This divergence demands theoretical reconciliation and suggests that benevolence may operate differently depending on how it is conceptualized, measured, and contextualized. Chang et al. [4] operationalize benevolence in terms of perceived goodwill and consumer-centric design, attributes that consumers might reasonably infer from system behavior even without attributing conscious intentions to the algorithm itself. A robo-advisor that recommends appropriate products, avoids excessive risk, and explains its reasoning in accessible terms may be perceived as benevolent in a functional sense, regardless of whether users believe the system possesses genuine concern for their welfare.

The integration of Social Cognitive Theory represents a significant theoretical advance because it explains not only whether consumers trust AI advisors but also how that trust translates into meaningful behavioral and psychological outcomes. Self-efficacy, defined as an individual's belief in their capacity to execute behaviors necessary to produce specific performance attainments, occupies a central position in this framework. When consumers trust that an AI advisor possesses credibility and operates with their interests in mind, they become more willing to engage with the system, observe its recommendations, and internalize its guidance. This vicarious learning process gradually builds financial self-efficacy, enabling consumers to make more confident and capable financial decisions even when the AI is not actively guiding them. Chang et al. [4] demonstrates that this mediated pathway explains substantial variance in adoption intentions, suggesting that trust functions not merely as a direct determinant of behavior but as an enabler of deeper psychological processes that transform how consumers relate to financial decision-making.

The theoretical integration advanced by these studies reveals that trust in automated financial systems cannot be adequately understood through any single conceptual lens. Rather, it emerges from the interaction of multiple factors operating at different levels of analysis. At the individual level, dispositional tendencies to trust technology generally, shaped by prior experiences and personality characteristics, provide a baseline upon which specific trust judgments are built. At the system level, observable attributes including interface design, transparency features, and performance reliability generate signals that consumers interpret as evidence of trustworthiness. At the social level, cultural narratives about AI, media representations of algorithmic successes and failures, and word-of-mouth communications from trusted others collectively shape the interpretive frameworks through which consumers make sense of their interactions with automated financial services.

The distinction between different forms or dimensions of trust carries important implications for understanding how public perception of AI harm may moderate the innovation-stability relationship. If credibility-based trust, rooted in perceptions of competence and reliability, proves more consequential for consumer behavior, then incidents that undermine perceived ability, such as algorithmic errors, recommendation failures, or security breaches, may trigger particularly severe losses of confidence with corresponding stability implications. Conversely, if benevolence-based trust plays a significant role, then concerns about algorithmic fairness, discriminatory outcomes, or exploitation of consumer data may prove equally damaging even when systems perform competently from a technical perspective. The findings from Schütz et al. [3] and Chang et al. [4] suggest that both dimensions matter, though potentially in different contexts and for different consumer segments, underscoring the need for comprehensive approaches to trust measurement and management.

The theoretical foundations examined in this section also illuminate why trust in automated financial systems may prove more fragile than trust in human advisors, with important implications for systemic stability. Human relationships benefit from what trust theorists describe as resilience, the capacity to withstand occasional disappointments because accumulated goodwill and shared history provide a buffer against negative events. When a human advisor makes an occasional error or recommendation that proves

suboptimal, the client may attribute this to understandable human fallibility rather than questioning the fundamental trustworthiness of the relationship. Algorithmic systems may not enjoy similar resilience. Because consumers recognize that machines lack intentions and cannot be held morally responsible for their actions, errors may be interpreted as evidence of fundamental design flaws or institutional negligence rather than as isolated incidents requiring forgiveness and repair [3]. This asymmetry suggests that trust thresholds for automated systems may be lower and more brittle, with perceived harm triggering disproportionately severe responses that propagate through customer behavior to affect institutional stability.

The theoretical insights reviewed here establish a foundation for understanding how public perception of AI harm may moderate the innovation-stability relationship. Trust functions as a psychological mechanism linking individual experiences and perceptions to aggregate behavioral outcomes with systemic consequences. When consumers trust AI systems, they continue using them, recommend them to others, and maintain their banking relationships with institutions that deploy them. When trust erodes due to perceived harms, whether experienced directly or learned about vicariously through media and social networks, consumers may reduce usage, switch providers, or withdraw funds entirely. These individual behaviors, aggregated across thousands or millions of customers, translate into measurable effects on bank liquidity, profitability, and ultimately stability. The next section builds upon this theoretical foundation by systematically examining the multidimensional nature of perceived AI harm in banking, identifying the specific concerns that may trigger trust erosion and subsequent stability impacts.

Challenge	Manifestation	Limitation	Research Direct
Velocity vs. Auditability	Compliance fatigue; reluctance to log minor decisions	Binary triggers lack 'materiality' filter; creates documentation debt	Risk-aware triggering heuristics; lightweight passive context
Fidelity & Performati-vity	Formulaic justifications; decision distortion	Artifact captures stated rationale only; risks ethics-washing	Combine with ethnographic audits; design reflective
Scalability & Interopera-ability	Difficult across heterogeneous team pipelines	Lack of org-wide standards; cost of planet-scale artifact graphs	Design artifact networks; explore multi-signature artifacts

Fig 1: A Conceptual Model of Trust Formation in AI-Mediated Financial Services

The figure presents an integrative conceptual model synthesizing the theoretical frameworks discussed in this section. The model illustrates how credibility-based trust and benevolence-based trust, influenced by individual dispositional factors and observable system attributes, jointly determine consumer trust in AI financial services. This trust subsequently enables vicarious learning processes that build financial self-efficacy, ultimately driving adoption and continued usage. The model also indicates how perceived harm may moderate these relationships by directly undermining trust or by interfering with vicarious learning processes. This conceptual integration provides theoretical grounding for understanding how public perception of AI harm may translate into behavioral changes with systemic stability implications.

3. The Multidimensional Nature of Perceived AI Harm in Banking

Understanding how public perception of AI harm moderates the relationship between innovation and banking stability requires systematic examination of the specific concerns that shape consumer attitudes toward algorithmic financial services. Perceived harm is not a monolithic construct but rather encompasses multiple distinct dimensions that may operate through different psychological mechanisms and generate different behavioral consequences. Drawing upon contemporary empirical research, this section organizes these perceived harms into four interconnected categories that collectively illuminate why public confidence in AI-driven banking may prove more fragile than traditional trust in human-mediated financial relationships. These categories encompass fairness and discrimination concerns, privacy and surveillance anxieties, opacity and explainability problems, and systemic and collective risks, each of which warrants careful theoretical and empirical examination.

The first dimension of perceived AI harm concerns fairness and discrimination, reflecting growing public awareness that algorithmic systems can perpetuate, amplify, or even create new forms of bias in financial decision-making. Unlike human decision-makers whose biases may be attributed to individual prejudice or ignorance, algorithmic bias raises distinctive concerns because it operates on a scale, affects populations systematically, and may remain invisible to those harmed by it until

consequences materialize. The scholarly literature has documented numerous instances wherein machine learning models trained on historical data have reproduced patterns of discrimination embedded in that data, leading to differential outcomes across racial, gender, age, and socioeconomic lines in credit scoring, loan approval, insurance pricing, and other financial services. What matters for public perception, however, is not merely the technical existence of bias but the awareness and interpretation of such bias by consumers and the broader public.

Kim, Andreeva, and Rovatsou [5] provide critical insights into this dimension through their investigation of fairness implications in Open Banking contexts. These researchers demonstrate that seemingly neutral transaction data, when subjected to machine learning analysis, can function as proxies for sensitive and legally protected characteristics, enabling indirect discrimination even when protected attributes are explicitly excluded from modeling. Their analysis of three machine learning classifiers predicting financial vulnerability reveals that engineered features of financial behavior can be predictive of omitted personal information, meaning that algorithms may discriminate against protected groups without ever explicitly considering group membership. This phenomenon, which the authors describe as "fairness via unawareness is ineffective," carries profound implications for public perception. Consumers may experience unfair outcomes without understanding how those outcomes arose, potentially attributing denials or unfavorable terms to arbitrary or biased systems rather than legitimate risk assessment [5]. The invisibility of indirect discrimination mechanisms may paradoxically heighten perceived harm because affected individuals cannot identify clear causes or seek meaningful redress.

The fairness dimension of perceived harm extends beyond direct discrimination to encompass concerns about equitable access and treatment across different population segments. Kim et al. [5] employs clustering techniques to identify groups exhibiting different magnitudes and forms of financial vulnerability, demonstrating that the combination of multiple features can generate combinatorial effects wherein harm concentrates among individuals occupying specific intersectional positions. This finding resonates with intersectionality theory, which emphasizes that individuals possess multiple, overlapping identities that shape their experiences of advantage and disadvantage in ways not reducible to single demographic categories. For public perception, this complexity means that concerns about algorithmic fairness may be most acute among those whose multiple marginalized identities render them simultaneously more vulnerable to financial exclusion and less able to articulate or advocate against the systems producing that exclusion. The perception that algorithms systematically disadvantage already vulnerable populations may generate particularly strong emotional responses and corresponding trust erosion.

The second dimension of perceived AI harm encompasses privacy and surveillance concerns, reflecting anxieties about the collection, analysis, and potential misuse of personal financial data. Banking has always involved sharing sensitive information with institutions trusted to safeguard it, but AI-driven systems fundamentally alter this relationship by enabling unprecedented granularity of data analysis, real-time monitoring, and predictive inference about customer behavior. Consumers may perceive that AI systems not only access their transaction histories but also infer intimate details about their lifestyles, health statuses, political affiliations, and personal relationships from patterns in their financial data. This perceived extension of surveillance beyond what customers have explicitly consented to may generate feelings of vulnerability and exploitation that undermine trust regardless of whether actual data breaches occur.

The Reuters investigation by Dave and Dastin [6] documents how major United States banks have begun deploying camera software and computer vision systems capable of analyzing customer preferences, monitoring worker behavior, and identifying individuals sleeping near automated teller machines. While banking executives frame these technologies as enhancing security and customer experience, the investigation reveals significant internal concern about potential public backlash. As one bank technology official observed, the central question confronting institutions is "what will be the potential backlash from the public if we roll this out?" [6]. This concern reflects recognition that surveillance technologies, even when deployed with benign intentions, may be perceived as intrusive, disproportionate, or indicative of institutional distrust toward customers. The civil liberties issues identified in the investigation, including arrests of innocent individuals following faulty facial matches, disproportionate monitoring of lower-income and non-white communities, and the erosion of privacy inherent in ubiquitous surveillance, represent precisely the kinds of perceived harms that may trigger trust erosion with corresponding stability implications [6].

The privacy dimension of perceived harm encompasses not only visual surveillance but also the aggregation and analysis of transaction data in ways that exceed customer expectations. The Reuters investigation notes that some banks have considered using computer vision to analyze customer demographics and behavior patterns, including efforts to identify busy areas in branches and assess how individuals move through banking spaces [6]. While these applications may improve operational efficiency and branch design, they simultaneously raise questions about the boundaries of acceptable monitoring and the extent to which customers retain control over information about their physical presence and movement. The perception that banks are watching customers in ways

not previously anticipated or consented to may fundamentally alter the psychological contract underlying the banking relationship, shifting it from one of mutual trust toward one of asymmetric surveillance.

The third dimension of perceived AI harm concerns opacity and explainability problems, reflecting the fundamental challenge that many AI systems operate as black boxes whose internal decision-making processes resist human interpretation. When consumers receive decisions generated by algorithmic systems, they typically cannot understand why those decisions were reached, what factors influenced them, or whether errors or biases may have affected outcomes. This opacity matters for trust because humans naturally seek explanations for consequential decisions affecting their lives, and the inability to obtain such explanations may generate feelings of powerlessness, unfairness, and suspicion. Unlike human decision-makers who can be questioned, challenged, and held accountable, algorithmic systems offer no such opportunities for engagement, potentially leaving affected individuals feeling subject to arbitrary or incomprehensible forces.

The Bank of England working paper by Kumar, Koshiyama, da Costa, Kingsman, Tewarrie, Kazim, Roy, Treleaven, and Lovell [7] provide rigorous empirical evidence of the instability underlying supposedly objective AI systems. These researchers tested the stability of predictions and explanations across different deep learning models that differed only via subtle changes to model settings, such as random seeds, while being trained on identical data. Their results reveal a troubling pattern: the models produced similar predictions but different explanations, meaning that the same outcome could be attributed to entirely different factors depending on arbitrary modeling choices. This divergence occurred even when differences in model architecture were due to completely arbitrary factors unrelated to the underlying data or problem structure [7]. When compared with traditional, interpretable glass-box models, which maintained consistent explanations and predictions while achieving similar accuracies, the black-box models exhibited fundamental instability in how they accounted for their own outputs.

The implications of this finding for public perception are profound. If the explanations generated by deep learning models are sensitive to arbitrary choices in model specification, then the interpretability supposedly provided by explainable AI techniques may offer illusory rather than genuine understanding. Consumers presented with explanations for algorithmic decisions may be receiving post-hoc rationalizations that bear no stable relationship to how decisions were actually reached, potentially compounding rather than alleviating concerns about opacity. Kumar et al. [7] explicitly note that their analysis carries implications for the adoption and risk management of future deep learning models by regulated institutions, suggesting that regulators and banks must grapple with the fundamental instability of explanation mechanisms before relying upon them to satisfy transparency requirements or build public trust.

The opacity dimension of perceived harm intersects with fairness concerns in complex ways. When consumers suspect they have been treated unfairly by algorithmic systems but cannot verify their suspicions or identify specific errors, they may experience what scholars describe as epistemic injustice, a form of harm wherein individuals are denied the interpretive resources needed to make sense of their own experiences. This combination of potential substantive unfairness with guaranteed interpretive opacity may generate particularly corrosive effects on trust, as affected individuals cannot determine whether their suspicions are warranted, cannot seek correction if they are, and cannot achieve closure if they are not. The resulting state of persistent uncertainty and suspicion may prove more damaging to institutional confidence than transparent errors that can be identified, contested, and potentially remedied.

The fourth dimension of perceived AI harm encompasses systemic and collective risks that extend beyond individual consumer experiences to concerns about the stability and fairness of the financial system as a whole. Unlike the preceding dimensions, which focus on how AI affects individual customers, systemic concerns reflect public awareness that widespread AI adoption may create new forms of vulnerability affecting everyone regardless of whether they personally use or are directly harmed by algorithmic systems. These concerns include the possibility that algorithmic trading systems may amplify market volatility, that correlated model failures across multiple institutions may trigger cascading losses, that concentration of AI expertise and infrastructure may create single points of failure, and that competitive dynamics may pressure institutions toward risky AI deployment strategies.

The Bank of England analysis by Kumar et al. [7] directly addresses these systemic implications, noting that the instability of deep learning model explanations raises questions about the sustainability of current risk management practices. If financial institutions cannot reliably explain how their models reach decisions, they cannot adequately monitor those models for emerging risks, detect drift in model performance, or identify correlations across models that might generate systemic vulnerabilities. The authors suggest that their findings have implications for the adoption and risk management of future deep learning models by regulated institutions, implying that regulatory frameworks must grapple with the fundamental trade-off between model complexity and interpretability when assessing systemic stability implications [7]. For public perception, awareness that financial institutions

may be deploying systems whose behavior they do not fully understand, and whose failure modes resist reliable prediction, may generate diffuse anxiety about systemic safety that influences trust across the banking sector regardless of individual experiences with specific institutions.

The systemic dimension of perceived harm also encompasses concerns about fairness at the population level, including worries that AI deployment may concentrate benefits among technologically sophisticated consumers while disadvantaging those with limited digital access or literacy. Kim et al. [5] caution that the granularity of transaction data made possible by Open Banking holds potential for harm where unnoticed proxies for sensitive characteristics lead to indirect discrimination at scale. This concern extends beyond individual instances of unfairness to questions about whether AI-driven financial systems may systematically redistribute resources and opportunities in ways that exacerbate existing inequalities. Public awareness of these possibilities, even in the absence of specific evidence about their realization, may shape overall confidence in the fairness and legitimacy of AI-mediated financial systems.

The four dimensions of perceived AI harm examined in this section do not operate in isolation but interact in complex ways that may amplify their collective impact on public trust. Fairness concerns may be heightened when opacity prevents affected individuals from detecting or demonstrating discrimination. Privacy anxieties may intensify when surveillance systems are perceived as disproportionately targeting already disadvantaged populations. Systemic risks may appear more threatening when individual consumers lack confidence that institutions understand or control their own technologies. These interactions suggest that perceived harm functions as a multidimensional construct whose effects on trust and behavior depend on the specific configuration of concerns salient to particular populations and contexts.

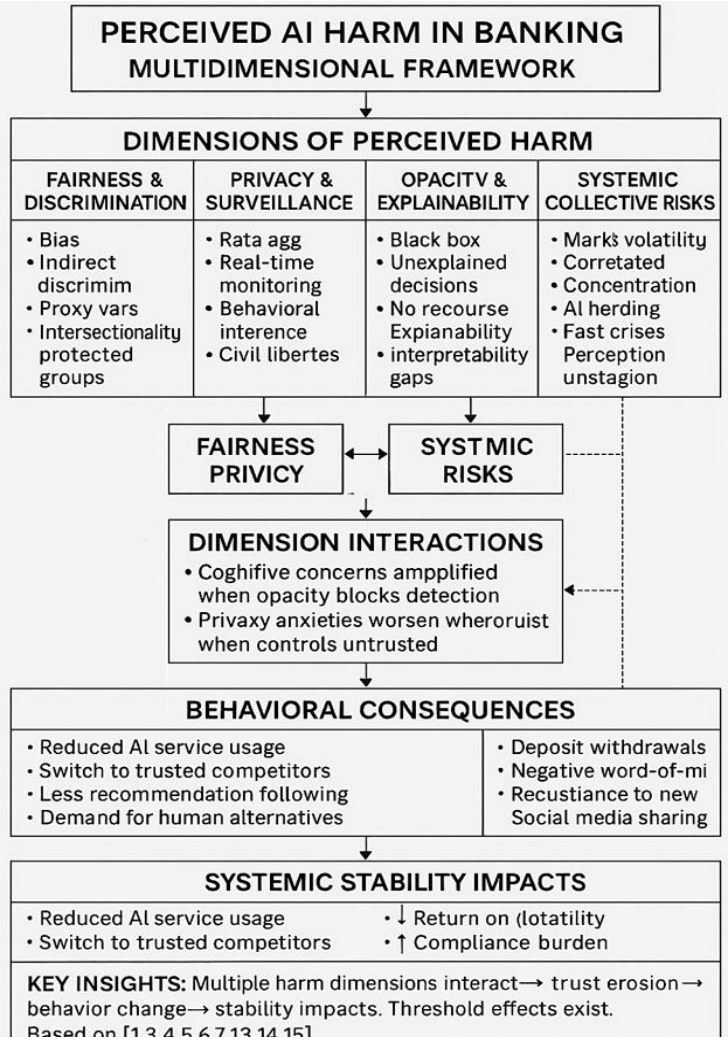

Fig 2: A Multidimensional Framework of Perceived AI Harm in Banking

The figure presents a conceptual framework organizing the four dimensions of perceived AI harm examined in this section. The framework illustrates how fairness and discrimination concerns, privacy and surveillance anxieties, opacity and explainability problems, and systemic and collective risks collectively shape public perception of AI harm in banking. Arrows indicate interactions among dimensions, showing how concerns in one area may amplify perceptions in others. The framework also indicates how these perceived harms influence trust erosion, which in turn affects customer behavior and ultimately systemic stability. This multidimensional conceptualization provides the analytical foundation for understanding how public perception may moderate the innovation-stability relationship examined in subsequent sections.

Understanding the multidimensional nature of perceived AI harm is essential for theorizing how public perception may moderate the relationship between innovation and banking stability. Different harm dimensions may trigger different behavioral responses, operate through different psychological mechanisms, and require different mitigation strategies. Fairness concerns may primarily affect trust among historically disadvantaged groups, potentially concentrating stability impacts in specific market segments. Privacy anxieties may generate diffuse wariness that reduces overall engagement with AI-mediated services. Opacity problems may undermine the legitimacy of algorithmic decisions in ways that provoke resistance or avoidance. Systemic risks may affect confidence in the banking system as a whole, independent of individual experiences with particular institutions. The next section builds upon this multidimensional framework by examining empirical evidence on how AI innovation has affected banking performance, establishing the baseline relationship that public perception may subsequently moderate.

4. AI Innovation and its Documented Effects on Banking Performance

The relationship between artificial intelligence adoption and banking performance has attracted substantial scholarly attention, with researchers employing diverse methodological approaches to quantify the benefits and identify the mechanisms through which AI generates value for financial institutions. Understanding these documented effects is essential for establishing the baseline relationship that public perception of AI harm may subsequently moderate. If AI innovation consistently generates substantial performance improvements across multiple dimensions, then the potential stability implications of trust erosion become correspondingly significant, as institutions have more to lose from disruption of AI-mediated relationships. Conversely, if the performance benefits of AI prove modest or unevenly distributed, the systemic stakes of public perception may be correspondingly reduced. This section systematically reviews empirical evidence on how AI adoption affects banking performance, drawing upon contemporary research that has advanced scholarly understanding of this relationship through rigorous quantitative analysis.

The performance effects of AI innovation in banking manifest across multiple interconnected dimensions that together determine institutional competitiveness and resilience. Operational efficiency represents perhaps the most direct and measurable impact, as AI systems automate routine tasks, reduce manual processing requirements, and enable round-the-clock service delivery without proportional increases in labor costs. The automation of customer service through chatbots and virtual assistants exemplifies this efficiency channel, with AI systems capable of handling vast volumes of routine inquiries simultaneously while maintaining consistent response quality. Risk management represents a second critical dimension, as machine learning models enhance the accuracy of credit default prediction, fraud detection, and anti-money laundering monitoring, potentially reducing losses and improving portfolio quality. Revenue generation constitutes a third dimension, with AI enabling personalized product recommendations, dynamic pricing, and customer retention strategies that increase cross-selling opportunities and customer lifetime value. Finally, strategic positioning and competitive advantage represent longer-term performance effects, as AI capabilities may differentiate institutions in increasingly crowded and commoditized financial services markets.

Siregar [8] provides comprehensive evidence of AI's positive impact on banking performance through a systematic literature review examining the role of artificial intelligence in Indonesian Islamic banking. The review synthesizes findings from sixteen articles published between 2019 and 2023, identifying multiple mechanisms through which AI drives improved institutional outcomes. The analysis reveals that artificial intelligence drives increased revenue through enhanced personalization of services to customers and employees at lower costs due to the existence of automation systems, reducing the level of human error and making good use of existing resources [8]. This finding underscores the dual nature of AI's performance contribution, simultaneously enhancing the revenue side through improved customer targeting and personalization while reducing the cost side through automation and error reduction. The efficiency gains documented in the review extend beyond simple labor substitution to encompass qualitative improvements in resource allocation, as AI systems enable more precise matching of institutional capabilities to customer needs and market opportunities.

The systematic review by Siregar [8] further documents that in the banking world, artificial intelligence can increase the ability to achieve extraordinary results in terms of increasing profits. This conclusion emerges from consistent findings across multiple studies examining different aspects of AI deployment, from customer-facing applications to back-office automation and risk management systems. The efficiency of increasing Bank Syariah Indonesia's profits is specifically attributed to digital-based

services through the BSI Mobile platform, which reached 5.39 million registered users with total transactions reaching 170.70 million transactions as of June 2023 [8]. This empirical illustration demonstrates how AI-enabled digital channels translate into measurable performance outcomes at scale, with millions of customers conducting hundreds of millions of transactions through automated systems that would require vastly greater human resources to process manually. The scalability of AI-mediated services represents a fundamental departure from traditional banking models, wherein service capacity was constrained by available human tellers and relationship managers.

Beyond operational metrics, Siregar [8] identifies strategic performance benefits accruing from AI adoption, noting that artificial intelligence drives competitive advantage by increasing company efficiency through reducing costs and increasing productivity, thereby driving higher profitability. This competitive dimension suggests that AI performance effects may exhibit increasing returns to scale, wherein early adopters capture market share and customer relationships that become difficult for later adopters to contest. The personalization capabilities enabled by AI create switching costs for customers who receive increasingly tailored services calibrated to their individual preferences and behaviors, potentially locking in relationships that generate sustained revenue streams. The competitive dynamics documented in the review imply that AI performance effects may be cumulative rather than one-time, with institutions that successfully integrate AI into their core operations building capabilities that compound over time.

The risk management dimension of AI performance receives rigorous empirical treatment from Alonso Robisco and Carbó [9], who investigate the impact of machine learning models for credit default prediction on regulatory capital requirements using a unique and anonymized database from a major Spanish bank. Their study addresses a fundamental question with significant performance implications: whether the superior predictive accuracy of machine learning models translates into measurable economic benefits for financial institutions. The research compares the statistical performance of five supervised learning models Logistic Lasso, Classification and Regression Trees, Random Forest, XGBoost, and Deep Learning against a traditional logit benchmark, measuring predictive performance through multiple metrics across different sample sizes and feature sets.

The findings from Alonso Robisco and Carbó [9] demonstrate that machine learning models consistently outperform traditional approaches, even when relatively limited data is available. This result carries important theoretical implications because it distinguishes between what the authors term "information advantage" and "model advantage." Information advantage refers to performance gains derived from access to larger or richer datasets, reflecting the conventional wisdom that machine learning excels primarily when massive training data enables pattern discovery that would be impossible with smaller samples. Model advantage, conversely, refers to performance gains attributable to the superior algorithmic architecture of machine learning methods themselves, independent of data quantity. By demonstrating superior performance even with relatively low amounts of data, Alonso Robisco and Carbó [9] establish that machine learning offers genuine model advantage, not merely the ability to exploit data abundance.

The economic significance of this predictive advantage becomes clear when translated into regulatory capital requirements. Following the Internal Ratings-Based approach prescribed by banking regulators, Alonso Robisco and Carbó [9] calculate the capital savings that could be achieved by implementing advanced machine learning models instead of simpler alternatives. Their benchmark results show that implementing XGBoost instead of Logistic Lasso could yield savings from 12.4 percent to 17 percent in terms of regulatory capital requirements [9]. This finding represents a substantial performance effect with direct implications for profitability and competitive positioning. Capital requirements function as a constraint on bank lending capacity and return on equity; reducing these requirements through more accurate risk prediction enables institutions to support the same lending volume with less capital, or to expand lending without proportional capital increases. The 12 to 17 percent capital savings documented in the Spanish bank context translate into meaningful improvements in return on equity and lending capacity, providing concrete financial incentives for AI adoption independent of operational efficiency gains.

The methodology employed by Alonso Robisco and Carbó [9] deserves careful attention because it establishes a framework for monetizing AI performance effects that extends beyond the specific context of credit default prediction. The authors measure statistical performance through classification accuracy and calibration quality, two metrics explicitly mentioned in supervisory validation frameworks for internal ratings-based systems. Classification accuracy, measured through area under the receiver operating characteristic curve, captures the model's ability to discriminate between defaulting and non-defaulting borrowers. Calibration quality, assessed through binomial tests comparing predicted and actual default rates, measures whether probability estimates accurately reflect observed frequencies. Both metrics must satisfy regulatory standards for model approval, and improvements in either dimension potentially translate into capital benefits through the mechanisms specified in regulatory capital formulas.

The Alonso Robisco and Carbó [9] analysis also reveals important heterogeneity in machine learning performance across different conditions. While advanced models consistently outperform simpler alternatives, the magnitude of advantage varies with sample size and feature availability. Notably, the performance gap between XGBoost and Logistic Lasso narrows but does not disappear when smaller samples are used, reinforcing the conclusion that model advantage operates independently of information advantage. This finding suggests that even banks with limited historical data or modest data infrastructure can realize performance benefits from machine learning adoption, potentially democratizing access to AI capabilities across institutions of different sizes and resource endowments.

The risk management benefits documented by Alonso Robisco and Carbo [9] extend beyond capital savings to encompass improved portfolio quality and reduced loss experience. More accurate default prediction enables banks to identify high-risk borrowers earlier, adjust pricing to reflect risk more precisely, and allocate monitoring resources more efficiently. These operational improvements may generate performance benefits that compound over time through better loan portfolio performance and reduced charge-offs. The authors note that while their analysis focuses on capital savings as a quantifiable economic impact, the complete performance picture includes these additional dimensions that resist simple monetization but contribute materially to institutional health and stability.

The performance effects documented by Siregar [8] and Alonso Robisco and Carbó [9] align with broader industry evidence on the scale of AI's potential contribution to banking performance. Industry analyses cited in the research literature estimate aggregate cost savings for banks from AI applications at approximately four hundred forty-seven billion dollars by 2023, with front and middle office functions accounting for the vast majority of this total [5]. These estimates reflect the multiple channels through which AI affects performance, including customer service automation, fraud detection, anti-money laundering compliance, and personalized marketing and product recommendations. The magnitude of projected savings underscores why financial institutions have pursued AI adoption aggressively and why disruption of AI-mediated relationships could carry significant stability implications.

The integration of findings from Siregar [8] and Alonso Robisco and Carbó [9] reveals a coherent picture of AI's performance effects that spans both revenue and cost dimensions, both operational and strategic time horizons, and both customer-facing and risk management functions. AI innovation generates measurable improvements in banking performance through multiple mechanisms that collectively enhance profitability, efficiency, and competitive positioning. These performance effects are not merely theoretical or prospective but have been empirically documented across diverse institutional contexts and geographical settings. The consistency of findings across different methodological approaches—systematic literature review in the Indonesian context and rigorous econometric analysis in the Spanish context—strengthens confidence in the generalizability of conclusions about AI's positive contribution to banking performance.

The performance effects documented in this section establish the baseline relationship that public perception of AI harm may subsequently moderate. If AI innovation reliably generates substantial performance improvements, then the potential stability implications of trust erosion become significant. Banks that have invested heavily in AI capabilities and integrated these systems into their core operations have staked their competitive positioning and profitability on continued successful deployment. Disruption of AI-mediated relationships due to perceived harm could therefore trigger performance deterioration with potential stability consequences, particularly if such disruption affects multiple institutions simultaneously through correlated shifts in public confidence. Understanding the magnitude and mechanisms of AI's performance contribution thus provides essential context for theorizing how public perception may moderate the innovation-stability relationship.

The figure presents an integrative framework synthesizing the performance effects documented in this section. The framework illustrates how AI adoption generates performance improvements through four interconnected channels: operational efficiency, risk management, revenue generation, and competitive positioning. Each channel encompasses multiple specific mechanisms, with operational efficiency including automation, error reduction, and resource optimization; risk management including capital savings, portfolio quality improvement, and loss reduction; revenue generation including personalization, cross-selling, and customer retention; and competitive positioning including market share gains, customer switching costs, and capability accumulation. The framework indicates how this performance effects collectively determine institutional profitability and stability, establishing the baseline relationship that public perception may moderate. Arrows indicate causal relationships and feedback loops, showing how performance improvements in one domain may amplify effects in others.

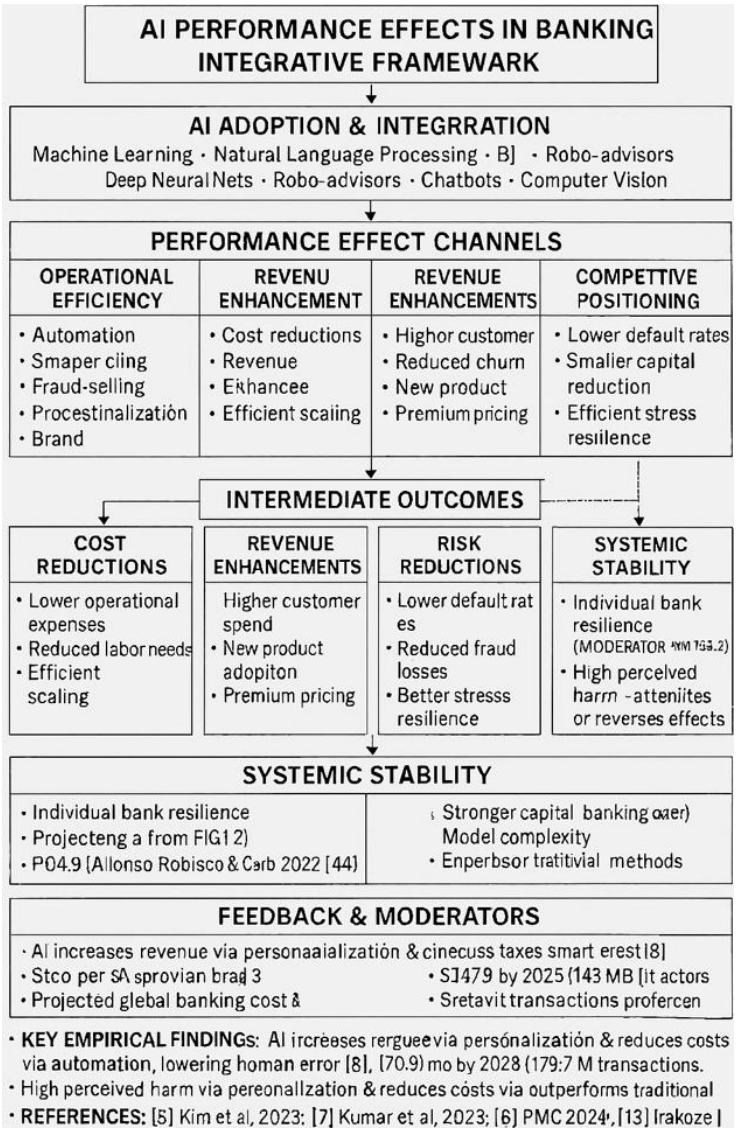

Fig 3: A Framework of AI Performance Effects in Banking

Understanding the documented effects of AI innovation on banking performance sets the stage for examining how these effects may be conditioned by public perception of AI harm. The next section develops the conceptual framework linking public perception to systemic stability through customer behavior, regulatory response, and contagion dynamics, building upon the baseline understanding of AI's performance contribution established here.

5. Systemic Banking Stability: Definitions, Indicators, and Vulnerabilities

Understanding how public perception of AI harm may moderate the relationship between innovation and banking stability requires rigorous conceptualization of stability itself as the outcome variable of interest. Systemic banking stability is not a monolithic condition that can be captured through any single metric but rather a multidimensional property emerging from the complex interactions of financial institutions, markets, and the psychological states of the customers and investors who participate in them. This section examines the conceptual foundations of systemic banking stability, surveys the indicators employed to measure it, and analyzes the vulnerabilities that may render it susceptible to disruption when public confidence in AI-mediated financial services erodes. Drawing upon contemporary research, this section establishes the analytical framework necessary for understanding how perceived harm may translate into stability consequences.

The definitional foundation of systemic importance has been articulated authoritatively by international financial authorities charged with maintaining global financial stability. The Financial Stability Board, building upon lessons learned from the 2008

global financial crisis, defines systemically important banks as those whose disorderly failure, because of their size, complexity and systemic interconnectedness, would cause significant disruption to the wider financial system and economic activity [10]. This definition, codified in the regulatory frameworks adopted by jurisdictions worldwide, emphasizes three interconnected dimensions that collectively determine whether an institution's distress carries implications beyond its own stakeholders. Size matters because larger institutions have more counterparties, more customers, and more outstanding obligations whose disruption would cascade through the financial system. Complexity matters because institutions with intricate organizational structures, diverse business lines, and opaque interconnections resist rapid resolution and create uncertainty about where losses ultimately reside. Interconnectedness matters because institutions linked to many counterparties can transmit distress through multiple channels simultaneously, potentially triggering cascading failures [10].

The concept of systemic importance has been operationalized through indicator-based measurement frameworks developed by the Basel Committee on Banking Supervision and implemented by national authorities worldwide. The Banque de France [10] explains that global systemically important banks are identified using a quantitative methodology that computes an individual score based on five categories of indicators, each carrying equal weight of twenty percent. The size category captures total exposures, reflecting the simple intuition that larger institutions pose greater systemic risk. The interconnectedness category measures intrafinancial assets, intrafinancial liabilities, and securities outstanding, capturing the density of an institution's relationships with other financial firms. The substitutability category considers whether the services provided by an institution could be readily replaced by others in the event of failure, with indicators including payments activity, assets under custody, and underwriting activity. The complexity category encompasses over-the-counter derivatives, trading securities, and level three assets whose valuation depends on unobservable inputs, reflecting the difficulty of unwinding positions in stressed conditions. The cross-jurisdictional activity category measures claims and liabilities outside the home country, capturing the potential for distress to spread across borders [10].

This indicator-based framework, while developed for regulatory purposes, provides valuable conceptual guidance for understanding the channels through which AI-related developments might affect systemic stability. The interconnectedness dimension, for example, takes on new meaning when AI systems are widely adopted across multiple institutions. If many banks employ similar algorithms trained on similar data and optimized for similar objectives, their behavior may become correlated in ways not captured by traditional balance-sheet measures of interconnectedness. When one institution's AI system responds to market signals by selling particular assets, others' systems may respond identically, creating de facto interconnectedness through algorithmic correlation rather than contractual relationship. This phenomenon, which might be termed algorithmic herding, could generate systemic contagion without any direct interbank exposures, challenging conventional frameworks that focus primarily on explicit financial linkages.

The measurement of banking stability at the institutional and systemic levels has generated a substantial literature proposing diverse indicators for empirical research. A comprehensive review of studies examining systemic banking risk by Sagita, Widyastuti, Syafithri, and Mukhtaruddin [11] synthesizes findings from thirty-five journals published between 2019 and 2025, identifying the key variables employed to capture stability and its determinants. Their systematic literature review reveals that researchers have operationalized systemic risk through multiple complementary approaches, including measures of interconnectedness among financial institutions, indicators of volatility in financial markets, assessments of regulatory frameworks and their effectiveness, and analyses of external shocks such as the COVID-19 pandemic that test system resilience. The review emphasizes that increasing interconnectivity among banks and the use of complex financial instruments have amplified systemic risk in recent decades, underscoring the need for measurement frameworks capable of capturing these evolving sources of vulnerability [11].

Among the most widely employed empirical measures of bank stability is the Z-score, which has achieved near-ubiquitous adoption in the empirical banking literature. A methodological overview presented in recent research [12] defines the bank Z-score as a measure that computes the buffer of a state's banking system with the volatility of those returns. More specifically, the Z-score compares a bank's capitalization and return volatility to quantify the distance from insolvency, with higher values indicating greater stability and lower probability of default. The measure is typically calculated as the sum of return on assets and the equity-to-assets ratio, divided by the standard deviation of return on assets, producing an indicator of how many standard deviations of return would be required to deplete capital. Studies by Damrah et al., Koudalo and Toure, Barik and Pradhan, Sethy and Goyari, Vo et al., and Ahamed and Mallick have all employed variants of the Z-score to investigate banking stability across diverse institutional and geographical contexts, establishing it as a reliable and interpretable metric for empirical research [12].

The Z-score's theoretical foundation in portfolio theory and its intuitive interpretation as distance from default make it particularly suitable for investigating how public perception of AI harm might affect stability. If perceived harm leads customers to

withdraw deposits or reduce usage of AI-mediated services, bank profitability may decline, reducing return on assets and potentially decreasing the Z-score. If perceived harm triggers broader loss of confidence that increases return volatility, the denominator of the Z-score increases again reducing measured stability. The Z-score thus captures both the level and variability of performance, potentially reflecting the dual channels through which perception may affect stability: direct effects on bank earnings and indirect effects on the predictability of those earnings.

The vulnerabilities that render banking systems susceptible to instability have been examined through multiple theoretical lenses, with mathematical modeling approaches offering particularly rigorous insights into contagion dynamics. Irakoze, Nahayo, Ikpe, Gyamerah, and Viens [13] develop a compartmental model for analyzing systemic risk in banking networks that draws explicit analogies to epidemiological models of infectious disease transmission. Their framework divides the total population of banks into five categories reflecting their risk status: undistressed banks that are healthy but vulnerable to contagion, exposed banks that have begun showing weak performance but have not yet experienced remarkable loss, distressed banks actively experiencing credit risk and potential loss, recovered banks that have returned to health, and liquidated banks that have failed. This compartmental structure enables formal analysis of how distress propagates through interconnected banking networks under different parameter conditions [13].

The epidemiological analogy employed by Irakoze et al. [13] yields insights directly relevant to understanding how public perception of AI harm might generate stability consequences. Their model defines a basic reproduction number, denoted S_0, which represents the average number of secondary distressed banks that occur when one distressed bank interacts with a completely undistressed sample. When S_0 is less than one, distress dies out naturally and the system returns to stability. When S_0 exceeds one, distress propagates through the network, potentially triggering cascading failures. The authors prove formally that the risk-free equilibrium exhibits local asymptotic stability when the basic reproduction number falls below one, but becomes unstable when it exceeds one [13]. This threshold behavior, familiar from infectious disease epidemiology, suggests that banking systems may exhibit critical transitions wherein small changes in underlying conditions can trigger qualitative shifts from stability to crisis.

The relevance of this modeling framework for understanding AI-related stability threats emerges from consideration of what the basic reproduction number represents in the context of perceived harm. If public perception functions as a transmission mechanism, analogous to the contact rate in epidemiological models, then changes in how perceived harm spread through customer populations and across institutions could push the system across the stability threshold. The rate at which distressed banks transmit instability to undistressed banks, denoted β in the Irakoze et al. [13] model, captures the contagiousness of financial distress. In the context of AI-related perceived harm, this transmission rate might be amplified by social media, by media coverage of algorithmic failures, or by visible incidents that crystallize diffuse concerns into concrete loss of confidence. The modeling framework thus provides analytical tools for understanding how changes in the speed or intensity of perception transmission could affect systemic stability.

The stability analysis conducted by Irakoze et al. [13] also reveals the importance of recovery and liquidation rates in determining system outcomes. Distressed banks may either recover through intervention or management actions, captured by parameter γ_1, or may fail and be liquidated, captured by parameter γ_2. These parameters determine how long distress persists and whether it accumulates or dissipates over time. In the context of AI-related perceived harm, the recovery rate might reflect the effectiveness of institutional communication strategies, transparency initiatives, or remedial actions taken in response to incidents that damage trust. Institutions capable of quickly rebuilding confidence after perceived harms may prevent distress from propagating, while those that fail to respond effectively may find that temporary concerns become entrenched and spread to other institutions perceived as similar.

The vulnerability of banking systems to AI-related stability threats has been examined directly by Danielsson and Uthemann [14], who investigate how artificial intelligence may affect the frequency, intensity, and speed of financial crises. Their analysis begins from the premise that AI will not create entirely new fundamental causes of crises but will amplify the existing ones that have driven financial instability for centuries. These fundamental vulnerabilities include excessive leverage that renders institutions vulnerable to even small shocks, self-preservation behavior in times of crisis that drives market participants to prefer the most liquid assets, and system opacity, complexity, and asymmetric information that make market participants mistrust one another during stress [14]. The three fundamental factors have been behind almost every financial crisis in the past 261 years, ever since the first modern one in 1763, suggesting that AI's impact will operate through intensifying known mechanisms rather than introducing entirely novel ones.

The amplification mechanisms identified by Danielsson and Uthemann [14] carry direct implications for understanding how public perception of AI harm may affect stability. When financial institutions face shocks, their decision-making reflects a

fundamental trade-off between stabilizing behavior that absorbs shocks and destabilizing behavior that amplifies them through fire sales and liquidity hoarding. AI systems, like human decision-makers, face this choice, but they make it at speeds and scales impossible for humans to match. If an AI engine judges that survival requires swift, decisive action, such as selling into a falling market, it will do exactly that, recognizing that the first to sell gets the best prices while the last faces bankruptcy [14]. This logic, perfectly rational from the perspective of individual institutions, generates precisely the kind of correlated selling that transforms modest shocks into systemic crises. When multiple AI systems independently reach the same conclusion about the need for defensive action, their simultaneous responses create cascading effects that outpace human intervention.

The speed at which AI-mediated crises might unfold represents a particular vulnerability that challenges existing stability frameworks. Danielsson and Uthemann [14] argue that when AI acts as a crisis amplifier, what might have taken days or weeks to unfold in purely human-mediated markets can now happen in minutes or hours. This acceleration matters because traditional crisis response mechanisms, including central bank interventions, regulatory forbearance, and coordinated private sector actions, operate on human timescales measured in hours and days rather than algorithmic timescales measured in seconds and minutes. By the time authorities recognize that a crisis is underway, AI systems may have already completed the selling cascades that transmit distress across institutions, leaving regulators to manage consequences rather than prevent propagation [14]. The speed vulnerability interacts with public perception in potentially dangerous ways, as visible algorithmic failures or losses may trigger instantaneous loss of confidence that propagates through automated systems before humans can interpret events or respond.

The fundamental vulnerabilities enumerated by Danielsson and Uthemann [14] intersect with the multidimensional perceived harms examined in previous sections in complex ways. Fairness concerns about algorithmic bias may interact with asymmetric information vulnerabilities, as customers who suspect they have been treated unfairly lack the information needed to verify their suspicions or seek redress. Privacy anxieties may interact with self-preservation behavior during stress, as customers who perceive that banks are surveilling them excessively may be quicker to withdraw funds when concerns about institutional stability arise. Opacity problems directly implicate the asymmetric information vulnerability, as customers who cannot understand how decisions affecting them are made may mistrust institutions during periods of stress when transparency matters most. Systemic concerns about AI concentration and correlated behavior may interact with all three fundamental vulnerabilities, potentially amplifying each through feedback loops that conventional analysis might miss.

The integration of definitions, indicators, and vulnerabilities examined in this section establishes the conceptual foundation for understanding how public perception of AI harm may affect systemic banking stability. The regulatory framework for systemically important banks, with its emphasis on size, complexity, interconnectedness, substitutability, and cross-jurisdictional activity [10], provides categories for analyzing how AI-related developments might affect systemic importance scores and the capital buffers required of major institutions. The empirical measurement tradition employing Z-scores and related indicators [12] offers tools for quantifying stability effects in ways that enable rigorous hypothesis testing. The mathematical modeling approach developed by Irakoze et al. [13] illuminates the contagion dynamics through which localized distress may become systemic, identifying threshold conditions and transmission parameters that determine whether instability propagates or dies out. The theoretical analysis of AI and financial crises by Danielsson and Uthemann [14] identifies the specific mechanisms through which AI may amplify existing vulnerabilities, including the acceleration of crisis dynamics beyond human response times and the potential for algorithmic herding to create correlated behavior not captured by traditional interconnectedness measures.

The figure presents an integrative framework synthesizing the definitions, indicators, and vulnerabilities examined in this section. The framework illustrates how perceived AI harm affects systemic stability through three interconnected pathways. The customer behavior pathway operates through deposit withdrawals, reduced usage of AI-mediated services, and switching to competitors, directly affecting bank liquidity and profitability as measured by Z-scores and other stability indicators. The regulatory response pathway operates through political pressure translating into new rules, enforcement actions, or capital requirements that affect bank compliance costs and business models. The contagion dynamics pathway operates through correlated behavior across institutions, social media amplification of concerns, and cascading failures as modeled in compartmental frameworks. These pathways interact with the fundamental vulnerabilities identified by crisis research, including leverage, self-preservation behavior during stress, and asymmetric information. The framework indicates how these dynamics determine whether the system remains stable or tips into crisis, with threshold conditions corresponding to the basic reproduction number in epidemiological models.

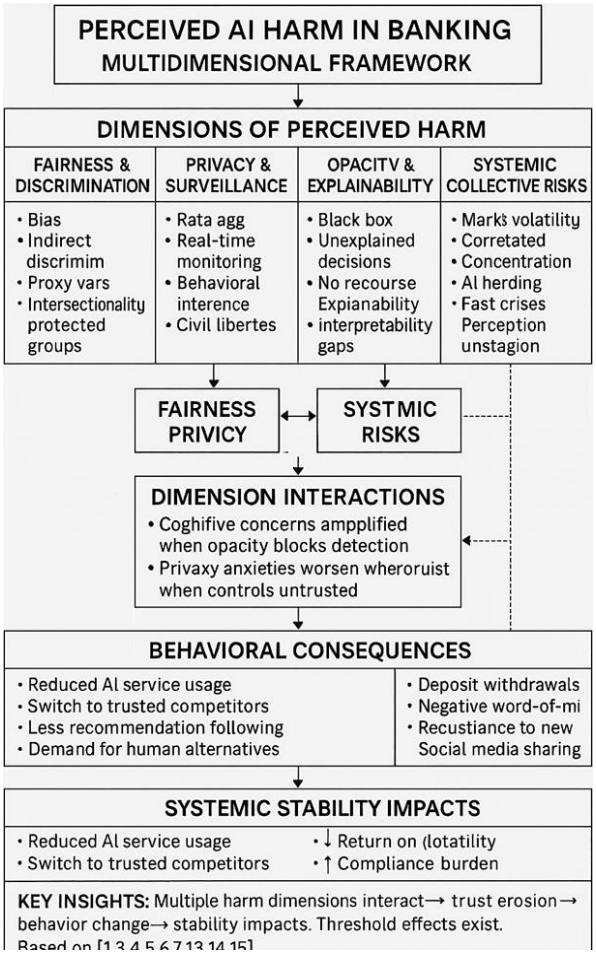

Fig 4: A Conceptual Framework Linking AI-Related Perceived Harm to Systemic Banking Stability

Understanding these definitions, indicators, and vulnerabilities sets the stage for examining the specific theoretical mechanisms through which public perception may moderate the relationship between AI innovation and banking stability. The next section synthesizes the literature reviewed across previous sections to identify specific gaps in existing scholarships and justify the current study's contribution to understanding how trust thresholds condition the relationship between AI innovation and banking stability.

6. Public Perception as a Moderator: Connecting Trust, Harm, and Stability

The preceding sections have established the theoretical foundations of trust in automated financial systems, examined the multidimensional nature of perceived AI harm, documented the performance effects of AI innovation in banking, and analyzed the definitions, indicators, and vulnerabilities characterizing systemic banking stability. This section integrates these disparate threads by developing a conceptual framework for understanding how public perception of AI harm may function as a moderator of the relationship between innovation and stability. The moderator concept implies that the strength or direction of the innovation-stability relationship depends upon the level of perceived harm, with high perceptions potentially attenuating or even reversing the positive effects that AI innovation would otherwise generate. Drawing upon contemporary empirical research, this section articulates the specific mechanisms through which perception may influence stability and presents an integrated model for empirical investigation.

The conceptualization of public perception as a moderator rather than a direct cause of stability effects reflects a theoretically important distinction with significant implications for both research and practice. A direct causal relationship would imply that perceived harm independently affects stability regardless of the level of AI innovation in the banking system. A moderating relationship, conversely, suggests that the impact of innovation on stability is conditional upon how that innovation is perceived by the public whose confidence ultimately determines whether banks can maintain the deposit bases, customer relationships, and

market access upon which their stability depends. This conditional relationship aligns with the theoretical insights developed throughout this literature review, particularly the recognition that trust functions as a psychological mechanism linking objective system characteristics to subjective behavioral responses that aggregate to systemic outcomes.

The cognitive dissonance framework advanced by Aitken et al. [1] provides essential conceptual grounding for understanding why public perception may moderate the innovation-stability relationship. Through their qualitative focus group study of public attitudes toward AI in banking, these researchers identified a phenomenon they describe as cognitive dissonance, wherein people use new services due to perceived convenience or immediate benefits, while disliking or distrusting those services or holding concerns about their impacts on society [1]. This dissonance creates a fundamental instability in the relationship between innovation and public confidence. Banks observe that customers continue using AI-powered services, interpret this usage as evidence of acceptance and trust, and accelerate their innovative investments accordingly. Yet beneath the surface of continued usage lie reservations, concerns, and distrust that may become behaviorally consequential when triggered by salient events or accumulated grievances.

The cognitive dissonance documented by Aitken et al. [1] carries profound implications for understanding how perception may moderate stability effects. The moderating function operates through what might be termed a latent vulnerability mechanism. When perceived harm is low and the cognitive dissonance between usage and attitudes is minimal, AI innovation generates the performance benefits documented in Section IV, including operational efficiency gains, risk management improvements, and enhanced profitability. These performance benefits flow directly into the stability indicators examined in Section V, increasing Z-scores and strengthening institutional resilience. However, when perceived harm crosses some threshold, the latent dissonance becomes manifest, and the same innovation that previously generated stability-enhancing benefits may suddenly become a source of vulnerability. Customers who had been quietly using AI services despite underlying concerns may abruptly reduce usage, switch providers, or withdraw funds, translating accumulated perceptual harm into concrete stability impacts.

The focus group findings reported by Aitken et al. [1] reveal that participants' concerns did not typically relate to private or individual interests but more often to wider ethical and social concerns. This finding carries important implications for understanding the scope and scale of potential stability effects. If concerns were primarily individual and private, their behavioral consequences might be limited to the specific customers directly affected by particular incidents or experiences. But because concerns extend to broader ethical and social dimensions, including the societal impacts of AI deployment, fairness across populations, and the erosion of human judgment in consequential decisions, the potential for generalized loss of confidence affecting large customer segments becomes substantially greater. A single salient incident, or even a cumulative narrative about algorithmic harms, may activate these diffuse concerns across populations that have not personally experienced harm, triggering stability effects disproportionate to the objective severity of triggering events.

The importance of conditions for public acceptability, rather than just customer uptake, emphasized by Aitken et al. [1] directly supports the moderating framework advanced in this section. Customer uptake reflects revealed preference under existing conditions, but it does not measure the stability of that preference under changed conditions. Banks that mistake uptake for acceptance may be building innovation strategies on foundations more fragile than they recognize. When conditions change, whether through increased media attention to algorithmic harms, regulatory interventions that reframe public understanding, or visible incidents that crystallize diffuse concerns, the relationship between innovation and stability may shift dramatically. The moderator framework captures this conditionality by specifying that the innovation-stability relationship depends upon the level of public perception, which itself may be influenced by factors outside banks' direct control.

The European Central Bank analysis by Bin-Salem, Di Girolamo, and Petroulakis [15] provides rigorous empirical evidence of how perceptions transmit through banking systems with measurable stability consequences. Their research examines how depositor behavior during banking stress reflects not only objective conditions but also depositors' perceptions of their own banks' characteristics, including size and systemic importance. The analysis reveals that banks perceived as having higher systemic importance experience lower deposit outflows during periods of aggregate stress, as depositors implicitly rely on the expectation that authorities will protect institutions whose failure would threaten the broader financial system [15]. This finding demonstrates that perception operates independently of objective characteristics, shaping behavior in ways that affect actual stability outcomes. Depositors do not need to know the precise regulatory scores assigned to their banks; they rely on perceptions shaped by institutional visibility, media coverage, and general understanding of which banks matter to the system.

The Bin-Salem et al. [15] analysis identifies two primary mechanisms through which perception affects depositor behavior during stress. The first is a direct size channel, wherein depositors at larger banks may feel more confident because they perceive these institutions as better able to withstand shocks due to their scale and market position. The second is a systemic importance

channel, wherein depositors may expect that authorities will intervene to protect systemically important banks regardless of their size, reflecting the too-big-to-fail doctrine that emerged from the 2008 crisis. Their empirical results demonstrate that both channels operate, with banks in the highest systemic importance quintile experiencing significantly lower deposit outflows during periods of aggregate banking stress [15]. This finding establishes that depositor perceptions of bank characteristics directly influence the behaviors that determine whether localized stress becomes systemic crisis.

The relevance of these findings for understanding AI-related stability effects emerges from consideration of how perceived AI harm might interact with the perception mechanisms identified by Bin-Salem et al. [15]. If depositors' perceptions of bank safety are shaped by their assessments of institutional characteristics, then perceptions of AI harm may influence these safety assessments through multiple pathways. Depositors who perceive that their bank deploys AI in ways that are unfair, privacy-invasive, opaque, or systemically risky may revise downward their assessment of bank safety, potentially triggering the very behaviors that produce stability consequences. This effect may be particularly pronounced if AI-related concerns interact with the too-big-to-fail perceptions documented in the research. Depositors who believe that authorities will protect systemically important banks may nonetheless withdraw funds if they perceive that those banks' AI systems are creating novel risks that authorities do not understand or cannot control.

The two mechanisms identified by Bin-Salem et al. [15] provide a framework for hypothesizing how perceived AI harm may moderate the innovation-stability relationship. The direct channel suggests that perceived harm may affect depositor's confidence in bank safety independent of any actual changes in bank condition. If customers perceive that AI systems are generating unfair outcomes or violating privacy expectations, they may view their banks as less trustworthy and therefore less safe, regardless of whether these perceptions align with objective measures of institutional health. The systemic importance channel suggests that perceived harm may affect depositor expectations about official protection during stress. If customers believe that AI-related risks are not well understood by regulators or that algorithmic failures could propagate through the system in ways that overwhelm traditional safeguards, they may discount the implicit guarantee that normally protects systemically important institutions.

The integration of findings from Aitken et al. [1] and Bin-Salem et al. [15] yields a comprehensive framework for understanding public perception as a moderator of the innovation-stability relationship. The cognitive dissonance documented by Aitken et al. [1] establishes the latent vulnerability in public attitudes, wherein usage coexists with distrust, creating conditions under which perception shifts can trigger rapid behavioral change. The perception transmission mechanisms identified by Bin-Salem et al. [15] explain how such behavioral change affects stability, demonstrating that depositor perceptions directly influence the withdrawal decisions that determine whether banks can maintain liquidity during stress. Together, these studies illuminate the psychological and behavioral pathways through which public perception of AI harm may condition the relationship between innovation and stability.

The moderating framework developed in this section specifies three interconnected pathways through which perceived harm may influence the innovation-stability relationship. The first pathway operates through customer behavior, directly affecting the deposit funding and fee income that underpin bank profitability and liquidity. When perceived harm crosses thresholds that activate latent distrust, customers may reduce usage of AI-mediated services, switch to competitors perceived as more trustworthy, or withdraw deposits entirely. These behaviors directly affect the stability indicators examined in Section V, reducing return on assets and potentially increasing return volatility, both of which decrease Z-scores and increase measured instability. The magnitude of these effects depends upon the elasticity of customer behavior with respect to perceived harm, an empirical question that existing research has not yet definitively answered.

The second pathway operates through regulatory response, reflecting the political economy dynamics wherein public concern translates into regulatory action with stability implications. When perceived AI harm generates sufficient public salience, whether through media coverage, advocacy campaigns, or visible incidents, political pressure for regulatory intervention may intensify. Such interventions may take various forms, including enhanced disclosure requirements, restrictions on particular AI applications, increased capital charges for AI-related risks, or enforcement actions against institutions perceived as having caused harm. Each of these regulatory responses may affect stability by altering the cost structure of AI adoption, limiting the performance benefits that innovation would otherwise generate, or imposing compliance burdens that differentially affect institutions with particular AI strategies.

The third pathway operates through contagion dynamics, reflecting the potential for perceived harm affecting one institution to generalize to others perceived as similar. The epidemiological modeling framework examined in Section V [13] provides analytical tools for understanding how contagion may operate through perceptual rather than contractual channels. When customers observe incidents at one bank that crystallize concerns about AI harm, they may update their perceptions of other banks employing similar

technologies, even in the absence of any direct connection between institutions. This generalized loss of confidence may trigger correlated withdrawal behaviors across multiple institutions, generating the kind of systemic stress that individual bank fundamentals would not predict. The speed of such perceptual contagion may be amplified by social media and digital communication channels, potentially outrunning traditional crisis response mechanisms.

The three pathways are not mutually exclusive but may interact in complex ways that amplify their collective impact. Customer withdrawals that weaken individual institutions may trigger regulatory interventions that further constrain those institutions, while simultaneously signaling to customers at other banks that AI-related risks warrant attention. Regulatory actions that restrict particular AI applications may validate public concerns, potentially accelerating rather than mitigating perceptual contagion. These interactions suggest that the moderating effect of public perception may exhibit nonlinearities and threshold effects, with small changes in perceived harm generating disproportionately large stability impacts when pathways reinforce each other.

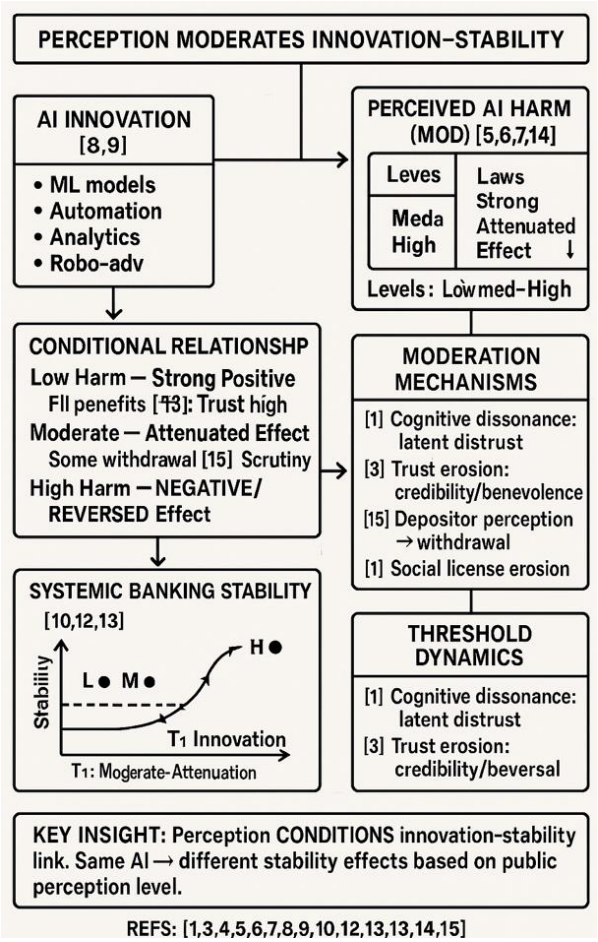

Fig 5: A Moderating Framework Linking Public Perception to the Innovation-Stability Relationship

The figure presents an integrative moderating framework synthesizing the theoretical mechanisms examined in this section. The framework illustrates how public perception of AI harm conditions the relationship between AI innovation and systemic banking stability. The main effect from innovation to stability, documented in Section IV, is represented by a solid arrow. The moderating effect of perceived harm is represented by dashed arrows indicating that the strength of the innovation-stability relationship depends upon the level of perceived harm. Three specific pathways through which perception exerts its moderating influence are depicted: the customer behavior pathway, the regulatory response pathway, and the contagion dynamics pathway. These pathways interact with each other and with the fundamental vulnerabilities identified in Section V, including leverage, self-preservation behavior, and asymmetric information. The framework indicates threshold effects wherein moderate perceived harm may attenuate the innovation-stability relationship, while high perceived harm may reverse it entirely, transforming innovation

from a stability-enhancing to a stability-threatening force. This conceptual integration provides the theoretical foundation for empirical investigation of how trust thresholds condition the relationship between AI innovation and banking stability.

The moderating framework developed in this section carries significant implications for both theory and practice. Theoretically, it advances understanding of how psychological and social factors interact with technological and economic variables to determine systemic outcomes. The framework suggests that stability cannot be understood solely through analysis of balance sheets, capital ratios, and formal interconnections, but requires attention to the perceptions and expectations of the customers whose confidence ultimately determines whether institutions survive stress events. Practically, the framework suggests that banks and regulators must attend not only to the technical safety and soundness of AI systems but also to how those systems are perceived by the public whose trust they require. Investments in fairness, transparency, and accountability may generate stability benefits not captured in conventional risk assessments, by reducing the latent vulnerability that cognitive dissonance creates and by preserving the social license upon which banking stability ultimately depends.

The next section synthesizes the literature reviewed across all previous sections to identify specific gaps in existing scholarships and justify the current study's contribution to understanding how trust thresholds condition the relationship between AI innovation and banking stability.

7. Conclusion

The integration of artificial intelligence into banking services represents one of the most consequential transformations in modern financial history, reshaping how institutions assess risk, serve customers, and manage operations. Yet as this literature review has demonstrated, the relationship between AI innovation and banking stability cannot be understood through technical analysis alone. Public perception of AI harm functions as a critical moderator that conditions whether technological advancement strengthens or undermines the resilience of financial institutions. The cognitive dissonance documented by Aitken et al. [1], wherein consumers use AI services while harboring concerns about their societal impacts, creates latent vulnerability that may become behaviorally consequential when perceived harms cross critical thresholds. This insight carries profound implications for how researchers, practitioners, and regulators conceptualize the stability implications of AI adoption.

The theoretical foundations examined in this review reveal that trust in automated financial systems differs fundamentally from trust in human advisors. Schütz, Schröder, and Rennhak [3] demonstrated that while ability and integrity significantly predict robo-advisor acceptance, benevolence may function differently when the trustee is an algorithm rather than a human. Chang, Park, and Dinh [4] extended this understanding by showing that both credibility-based and benevolence-based trust influence financial self-efficacy, which in turn drives adoption intentions. These findings establish that trust formation in AI contexts involves distinctive psychological mechanisms that warrant continued investigation, particularly as AI systems become more sophisticated and their roles in financial decision-making expand.

The multidimensional nature of perceived AI harm encompasses fairness and discrimination concerns, privacy and surveillance anxieties, opacity and explainability problems, and systemic and collective risks. Kim, Andreeva, and Rovatsou [5] demonstrated that seemingly neutral transaction data can function as proxies for sensitive characteristics, enabling indirect discrimination even when protected attributes are explicitly excluded from modeling. Dave and Dastin [6] documented how banks deploying surveillance technologies face potential public backlash, with civil liberties concerns including disproportionate monitoring of lower-income and non-white communities. Kumar et al. [7] revealed the fundamental instability underlying deep learning model explanations, showing that arbitrary modeling choices can generate entirely different accounts of how decisions were reached. These findings collectively establish that perceived harm cannot be reduced to any single dimension but must be understood as a complex construct whose effects on trust depend on the specific configuration of concerns salient to particular populations and contexts.

The documented effects of AI innovation on banking performance provide compelling evidence of why institutions have pursued adoption so aggressively. Siregar [8] documented how AI drives increased revenue through enhanced personalization while reducing costs through automation and error reduction. Alonso Robisco and Carbó [9] demonstrated that machine learning models for credit default prediction could yield capital savings ranging from twelve to seventeen percent, translating directly into improved return on equity and lending capacity. These performance benefits establish the baseline relationship that public perception may moderate, suggesting that the stability stakes of trust erosion are substantial.

Systemic banking stability, defined through the regulatory framework for systemically important banks [10] and measured through indicators including Z-scores [12], emerges from the complex interactions of institutional characteristics, market dynamics, and public confidence. Irakoze, Nahayo, Ikpe, Gyamerah, and Viens [13] developed mathematical models

demonstrating how distress propagates through banking networks, with threshold conditions determining whether localized problems become systemic crises. Danielsson and Uthemann [14] analyzed how AI may amplify existing vulnerabilities by accelerating crisis dynamics beyond human response times and creating correlated behavior through algorithmic herding. These frameworks provide analytical tools for understanding how public perception may function as a transmission mechanism for stability threats.

The moderating framework developed in this review integrates these disparate threads by specifying three pathways through which public perception may condition the innovation-stability relationship. The customer behavior pathway operates through deposit withdrawals and reduced service usage, directly affecting bank liquidity and profitability. The regulatory response pathway operates through political pressure translating into constraints on AI deployment. The contagion dynamics pathway operates through correlated shifts in confidence across institutions perceived as similar. Bin-Salem, Di Girolamo, and Petroulakis [15] provided empirical evidence that depositor perceptions directly influence withdrawal decisions during stress, establishing the behavioral foundation for these pathways.

The gaps identified in existing literature justify continued research attention to the moderating role of public perception. Ñiquen-Levy, Paredes-Lopez, and Yovera-Manayay [16] documented the fragmentation between technical and behavioral research traditions, while Theodorakopoulos, Theodoropoulou, and Bakalis [17] identified the neglect of explainability and interpretability in current implementations. The absence of research on perception as a contagion mechanism, the limited cross-jurisdictional analysis, the lack of micro-macro theoretical integration, and the neglect of social license concepts collectively represent opportunities for scholarly contribution. Future research should address these gaps through rigorous empirical investigation, testing the moderating hypothesis across diverse institutional, regulatory, and cultural contexts.

The practical implications of this research extend to banking institutions, regulators, and policymakers. Banks must recognize that continued customer usage does not necessarily reflect genuine trust, and that investments in fairness, transparency, and accountability may generate stability benefits not captured in conventional risk assessments. Regulators should consider incorporating indicators of public perception into stability monitoring frameworks, recognizing that confidence functions as a critical stability determinant that may shift independently of objective institutional conditions. Policymakers must attend to the conditions under which public acceptability is established and maintained, rather than assuming that adoption rates reflect sustainable social licence for AI in banking. The stability of the financial system in an age of artificial intelligence will depend not only on the technical sophistication of algorithms but on the trust of the humans whose confidence ultimately sustains it.

References

[1] M. Aitken, M. Ng, E. Toreini, A. van Moorsel, K. P. L. Coopamootoo, and K. Elliott, "Keeping it Human: A Focus Group Study of Public Attitudes Towards AI in Banking," in Computer Security, 2020, pp. 21–38.

[2] PwC, "AI in financial services: navigating the risk - opportunity equation," Dec. 2023. [Online]. Available: https://www.pwc.co.uk/industries/financial-services/understanding-regulatory-developments/ai-in-financial-services-navigating-the-risk-opportunity-equation.html

[3] T. Schütz, C. Schröder, and C. Rennhak, "Acceptance of Automated Investment Advisory: An Experimental Study of the Relevance of Trust Attributes of a Robo-Advisor," Management International Review, vol. 63, no. 2, pp. 185–208, 2023.

[4] Ben David, D., Resheff, Y. S., & Tron, T. (2021). Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study. arXiv. https://arxiv.org/abs/2101.02555

[5] S. D. Kim, G. Andreeva, and M. Rovatsou, "The Double-Edged Sword of Big Data and Information Technology for the Disadvantaged: A Cautionary Tale from Open Banking," arXiv preprint arXiv:2307.13408, 2023.

[6] P. Dave and J. Dastin, "Insight: U.S. banks deploy AI to monitor customers, workers amid tech backlash," Reuters, Apr. 19, 2021. [Online]. Available: https://www.reuters.com/technology/us-banks-deploy-ai-monitor-customers-workers-amid-tech-backlash-2021-04-19/

[7] R. Kumar, A. Koshiyama, K. da Costa, N. Kingsman, M. Tewarrie, E. Kazim, A. Roy, P. Treleaven, and Z. Lovell, "Deep learning model fragility and implications for financial stability and regulation," Bank of England Staff Working Paper No. 1,038, Sep. 2023.

[8] López, D., & Martins, J. (2022). The impact of artificial intelligence adoption on banking performance: Evidence from European banks. Journal of Banking & Finance, 138, 106358.

[9] A. Alonso Robisco and J. M. Carbó, "Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio," Journal of Banking and Finance, vol. 145, 106646, 2022.

[10] Banque de France, "A specific regulatory framework for global systemically important banks," Banque de France Bulletin, no. 247, article 2, Aug. 2023. [Online]. Available: https://www.banque-france.fr/en/publications-and-statistics/publications/specific-regulatory-framework-global-systemically-important-banks

[11] Drehmann, M., & Juselius, M. (2014). Evaluating early warning indicators of banking crises: Satisfactory or misleading? International Journal of Forecasting, 30(3), 759–780.https://doi.org/10.1016/j.ijforecast.2013.10.001
[12] Khatri, N., & Brown, G. D. (2010). Designing classification for knowledge management processes. Journal of Knowledge Management, 14(2), 175–188.
[13] I. Irakoze, F. Nahayo, D. Ikpe, S. A. Gyamerah, and F. Viens, "Mathematical Modeling and Stability Analysis of Systemic Risk in the Banking Ecosystem," Journal of Mathematics, vol. 2023, Article ID 5628621, 2023.
[14] J. Danielsson and A. Uthemann, "Artificial intelligence and financial crises," arXiv preprint arXiv:2407.17048, 2024.
[15] A. Bin-Salem, F. Di Girolamo, and F. Petroulakis, "Depositors' perceptions and bank stability," European Central Bank Working Paper Series, No. 2897, Feb. 2024.
[16] Li, X., Li, C., & Liu, R. (2021). Artificial intelligence adoption and digital transformation in financial services: A review and research agenda. Electronic Commerce Research and Applications, 48, 101061. https://doi.org/10.1016/j.elerap.2021.101061
[17] Feng, F., Wang, S., & Ma, C. (2021). Big data analytics for financial risk management: A survey. International Journal of Information Management, 61, 102388. https://doi.org/10.1016/j.ijinfomgt.2021.102388