The Economics of Artificial Intelligence Trust: Calibration as Public Good, Auditing as Market Failure, and the Governance of AI as a Credence Good

Nicholas Ward

Abstract

The integration of artificial intelligence into economic life has created a distinctive economic challenge that existing market and regulatory frameworks are poorly equipped to address: the systematic underproduction of calibrated AI trust. Artificial intelligence systems, particularly large language models, function as credence goods—products whose quality consumers cannot evaluate even after consumption—creating information asymmetries that prevent markets from efficiently pricing AI reliability. The calibration of AI systems—the accurate representation of their own knowledge boundaries—constitutes a public good that is structurally underproduced because individual AI developers cannot capture the social returns on calibration investment. The auditing infrastructure necessary to verify AI reliability faces compounding market failures that the Verification Tax analysis reveals are inherent rather than incidental. Drawing on three foundational references and twelve supplementary citations spanning the economics of trust, credence goods markets, public goods theory, institutional economics, and AI regulation, this paper develops a comprehensive economic framework for understanding and addressing the AI trust deficit. The analysis reveals that the MIRROR benchmark's documentation of systematic AI miscalibration, the Verification Tax's exposure of structural auditing constraints, and the practical demand for explainability in human resource analytics collectively illuminate an economic phenomenon: the AI trust market is failing in ways that require coordinated institutional intervention. The paper concludes by proposing an AI Trust Infrastructure Framework that combines market mechanisms, regulatory standards, and institutional innovations to address the fundamental economic dimensions of the AI trust challenge.

Keywords: Economics of AI Trust, Credence Goods, Public Goods, AI Calibration Market, Verification Market Failure, Institutional Trust, AI Governance Economics, Trust as Public Good, Information Asymmetry, AI Regulation

1. Introduction

Every economic transaction involves trust: the willingness of one party to accept exposure to the actions of another on the basis of expectations about those actions, without the ability to fully verify those expectations ex ante. Markets have developed elaborate institutions—certification, warranties, reputation mechanisms, regulatory frameworks—to address the trust challenges that arise in ordinary commercial transactions. Yet the deployment of artificial intelligence in consequential economic decisions has introduced a new and qualitatively distinct trust challenge for which existing market and regulatory institutions are poorly prepared.

Artificial intelligence systems, particularly the large language models that increasingly inform organizational and individual decisions, exhibit a property that distinguishes them from most goods and services traded in markets: their quality—the reliability of the knowledge and recommendations they produce—is fundamentally unknowable to consumers even after the moment of consumption. An AI system that provides a confident but incorrect medical diagnosis, a flawed strategic analysis, or a biased hiring recommendation inflicts harm that cannot be detected at the moment of decision, nor easily traced to its algorithmic source after the fact. Economists call goods with this property credence goods—products whose value consumers cannot accurately assess even with post-purchase experience.

The economic implications of AI's credence good character are profound. When consumers cannot evaluate product quality, markets fail to reward quality and punish poor quality with the efficiency that characterizes markets for ordinary goods. Producers of AI systems have weak incentives to invest in the accuracy and reliability that would justify consumer trust, because the mechanisms through which markets normally discipline quality—customer defection, reputation loss, legal liability—are systematically weakened. The result is a structural tendency toward the underproduction of AI reliability that no amount of competitive pressure alone can correct.

This economic analysis is grounded in three foundational references. Paper 1, Wang (2026), introduced the MIRROR benchmark for metacognitive calibration in large language models, documenting that even state-of-the-art AI systems systematically misrepresent their own reliability. The economic implication is that the calibration of AI systems—their capacity to accurately communicate what they know and do not know—is being produced at levels that are far below what social welfare would demand. Paper 2, Wang (2026), analyzed the Verification Tax—the resource constraints that limit comprehensive AI auditing—revealing that the auditing market for AI reliability faces structural challenges that prevent efficient verification even when strong incentives for verification exist. Paper 3, Bei et al. (2025), demonstrated the practical demand for explainable AI in strategic human resource analytics, illustrating how organizational buyers of AI services experience the consequences of the AI trust deficit in concrete, high-stakes decisions about people's careers.

The paper proceeds as follows. Section 2 develops the economics of AI as a credence good, drawing on the theoretical literature on credence goods markets and their well-documented tendencies toward market failure. Section 3 analyzes calibration as a public good, examining why individual AI developers will systematically underinvest in calibration relative to social optima. Section 4 examines the market for AI auditing and the verification market failures that compound the credence goods and public goods problems. Section 5 integrates the MIRROR and Verification Tax findings into the economic framework. Section 6 presents the AI Trust Infrastructure Framework as a comprehensive approach to addressing the AI trust market failure. Section 7 concludes with implications for policy and future research.

---

2. The Economics of AI as a Credence Good

2.1 Credence Goods: Theory and Evidence

Credence goods are products whose quality consumers cannot evaluate even after purchase and consumption. The classic examples—medical services, auto repair, legal advice—share a common structure: the consumer cannot independently determine whether the good's properties match their needs, and even experienced consumers often cannot assess whether the service they received was necessary, adequate, or optimal. The economics of credence goods markets, developed by Emons, Dulleck, and others, document a consistent pattern: credence goods

markets are characterized by systematic quality problems because producers face weak market discipline for poor quality.

The key insight from the credence goods literature is that information asymmetry—the producer knows more about product quality than the consumer—is not merely a description of the market but a cause of its failure. When consumers cannot evaluate quality, they cannot reward quality with repeat purchases or punish poor quality with defection. Producers of credence goods therefore face a distorted incentive structure in which investment in quality is inadequately rewarded and the temptation to reduce quality while maintaining appearance of adequacy is enhanced.

The credence goods literature identifies several mechanisms that partially address market failure in these contexts. Professional certification—through licensing, accreditation, or credentialing—attempts to substitute for consumer evaluation by providing third-party verification of producer quality. Warranties and guarantees provide post-consumption signals of producer confidence in quality. Reputation mechanisms—particularly in repeated transaction settings—create incentives for quality maintenance even in the absence of direct quality evaluation. Yet each of these mechanisms has limitations that are especially acute for AI systems.

2.2 AI as the Paradigmatic Credence Good

Artificial intelligence systems exhibit the properties of credence goods in an especially pronounced form. The quality of an AI system's outputs—the accuracy of its predictions, the reliability of its recommendations, the calibration of its confidence estimates—cannot be directly evaluated by most consumers even after the AI has been used. A clinical AI system that recommends a treatment may be right or wrong, but the physician who receives the recommendation typically cannot independently verify the recommendation's correctness before deciding whether to follow it. A strategic AI system that advises against an acquisition may be protecting the organization from a bad deal or may be protecting organizational inertia against a beneficial opportunity.

Several properties of AI systems make the credence goods problem especially severe. First, the complexity of AI reasoning makes quality evaluation requiring specialized expertise that most consumers do not possess and cannot afford to acquire. Second, the consequentiality of AI-influenced decisions—the domains in which AI is deployed, from healthcare to finance to employment, are precisely those where errors are costly and often irreversible. Third, the pace of AI development means that even experts have difficulty keeping pace with the capabilities and limitations of cutting-edge systems. Fourth, the opaqueness of large neural networks means that even the AI developers themselves often cannot explain why the system produced a particular output—making quality guarantees from developers themselves unreliable.

Research on trust, trustworthiness, and AI governance demonstrates that the information asymmetry in AI markets extends beyond product quality to fundamental questions about AI behavior. As AI systems become more agentic—capable of taking actions, adapting to contexts, and pursuing objectives in ways that are not fully predictable—the trust challenge shifts from evaluating the quality of a known product to evaluating the reliability of an entity whose behavior is inherently uncertain.

2.3 The Market for AI Trust: Demand Without Supply

The credence goods problem in AI markets creates a distinctive economic dynamic: demand for trust exists but cannot be efficiently satisfied by market supply. Consumers—organizations and individuals making decisions with AI assistance—want reliable AI systems, but their ability to distinguish reliable from unreliable AI is fundamentally limited. This demand, unable to be

converted into market discipline on AI producers, creates a structural gap between what consumers want and what producers have incentive to provide.

Akers and others have documented how markets for credence goods develop institutional responses to information asymmetry. Professional certification—through industry standards, third-party testing, and regulatory requirements—represents an attempt to substitute collective evaluation for individual consumer evaluation. Warranties and service agreements attempt to align producer incentives with consumer interests through contractual commitments. Yet each of these mechanisms faces specific challenges in AI markets. Industry standards for AI quality are nascent and contested. Third-party testing is hindered by the Verification Tax constraints on comprehensive AI evaluation. Contractual warranties for AI performance are limited by the difficulty of specifying and measuring AI quality in contractually meaningful terms.

---

3. Calibration as a Public Good: The Underinvestment Problem

3.1 Calibration as a Non-Excludable, Non-Rivalrous Good

The calibration of AI systems—their capacity to accurately represent what they know and do not know—possesses the defining properties of a public good: it is non-excludable (once an AI system is calibrated, all users benefit from its accurate uncertainty communication regardless of whether they contributed to the calibration effort) and non-rivalrous (one user's benefit from AI calibration does not reduce another user's benefit). These properties create the conditions for a classic public goods problem: the social returns to AI calibration investment exceed the private returns that individual AI developers can capture, leading to systematic underinvestment relative to the socially optimal level.

The public goods character of AI calibration has several dimensions. At the most basic level, calibration is a property of the AI system itself: a calibrated model accurately represents its own reliability to all users simultaneously. More broadly, the development of calibration methods—through research on benchmark frameworks like MIRROR, training techniques that improve calibration, and evaluation standards that measure calibration—constitutes a public good whose benefits accrue to the entire ecosystem of AI users.

The economics of public goods predict that when private actors cannot capture the social returns on investment, they will underinvest relative to the social optimum. For AI calibration, this prediction is strongly confirmed by the MIRROR benchmark findings: state-of-the-art AI systems, developed by sophisticated actors with substantial resources, exhibit systematic miscalibration. This is not because AI developers are incompetent or indifferent—it is because the incentive structure they face does not reward calibration investment commensurate with its social value.

3.2 Why Calibration Markets Fail

The market failure in AI calibration arises from several reinforcing mechanisms. First, the difficulty of measuring calibration—documented by the MIRROR benchmark's need for sophisticated multi-dimensional evaluation frameworks—means that even well-intentioned AI developers lack the metrics necessary to demonstrate the calibration quality of their systems. Without measurable quality, competitive markets cannot reward calibration. Second, the competition among AI developers emphasizes accuracy and capability rather than calibration, creating incentive structures that reward capability investment while leaving calibration investment undersupplied. Third, the information asymmetry between AI developers and users means that users cannot

distinguish calibrated from miscalibrated systems, eliminating the market mechanism through which consumer choice would otherwise discipline quality.

Arrow's foundational observation that trust-based unwritten guarantees are preconditions for trade and production acquires new significance in the AI context. The trust that economic actors place in AI systems—when that trust is misplaced due to AI miscalibration—is not merely a technical failure but an economic one: it undermines the transactions, investments, and organizational decisions that depend on AI-generated knowledge. The social cost of AI miscalibration thus extends far beyond the direct errors that miscalibrated AI produces to include the broader economic costs of degraded trust in the knowledge systems on which modern economies depend.

3.3 The Economic Consequences of Misaligned Incentives

The incentive misalignment in AI calibration has observable economic consequences. Organizations that deploy miscalibrated AI systems make systematically biased decisions: they over-rely on AI recommendations when those recommendations are confidently wrong and under-rely on AI recommendations when those recommendations are correctly uncertain. The resulting decision quality degradation represents a deadweight loss that is not reflected in any market transaction but is borne by organizations, individuals, and the broader economy.

The literature on trust and AI in Electronic Markets demonstrates that trust mechanisms are especially critical for AI adoption, yet the development of these trust mechanisms—calibration standards, auditing frameworks, reputation systems—is itself subject to the same public goods dynamics that undermine AI calibration itself. The market for AI trust thus faces a double public goods problem: the underlying good (calibrated AI) is underproduced, and the mechanisms for verifying and communicating that good (auditing, certification, reputation) are also underproduced due to incentive misalignment.

---

4. The Market for AI Auditing: Verification Market Failures

4.1 The Verification Tax as Economic Analysis

Paper 2's Verification Tax analysis provides an economic diagnosis of the AI auditing market failure. The Verification Tax identifies three dimensions of cost that prevent comprehensive AI auditing: statistical costs (the need for large test datasets, particularly for rare-event evaluation), mechanistic costs (the difficulty of tracing causal pathways through complex models), and institutional costs (the shortage of trained auditors and regulatory infrastructure). From an economic perspective, each of these dimensions represents a distinct source of market failure that prevents the efficient production and certification of AI reliability.

Statistical costs create a classic problem of adverse selection: the users most likely to need comprehensive AI auditing—those deploying AI in high-consequence, high-variability applications—are precisely those for whom comprehensive auditing is most expensive. The auditing market therefore serves the lower-stakes, lower-variability deployments while leaving the highest-risk deployments most underserved. This is precisely the adverse selection pattern that makes insurance markets fail: the risky pools are the ones that most need coverage but can least afford it.

Mechanistic costs create a problem of measurability that undermines the basic premise of quality certification. If the quality of an AI system cannot be feasibly measured, then the certification market cannot perform its function of credibly distinguishing high-quality from low-quality systems. The result is a market in which certifications proliferate but provide limited information about actual system quality—undermining the trust that certifications are supposed to create.

Institutional costs represent a coordination failure: the social value of a robust AI auditing infrastructure—a trained auditor workforce, standardized evaluation protocols, regulatory frameworks—is far greater than any individual actor's incentive to invest in creating it. This is the classic logic of public goods applied to institutional infrastructure: the returns to auditing infrastructure are non-excludable and non-rivalrous, leading to systematic underinvestment relative to social optima.

4.2 Why Markets Cannot Solve the Verification Problem

The market failure in AI auditing is not merely a matter of scale—more investment would solve it. The Verification Tax analysis implies that the fundamental constraints on AI auditing are inherent: they arise from the structural properties of complex AI systems and the statistical realities of rare-event evaluation. This means that even well-functioning markets with strong demand for AI reliability will produce insufficient auditing—because the product that the market is trying to produce (verified AI reliability) is inherently difficult to produce at scale.

The economics of trust and AI governance demonstrates that trust in AI systems is influenced by both institutional trust (trust in the organizations and frameworks that govern AI) and interpersonal trust (trust in the technology itself). Markets that substitute institutional trust for direct evaluation of AI quality can partially address the credence goods problem, but only to the extent that the institutions themselves can be trusted—which requires those institutions to have access to the verification information that the Verification Tax analysis reveals is fundamentally limited.

Recent research on trust formation and repair in human-AI financial advisory found that the dynamics of AI trust follow distinctive patterns that deviate from those observed in human-to-human trust. Critically, mere knowledge that advice originates from AI leads to overreliance, causing users to follow AI recommendations even when those recommendations contradict available contextual information. This finding has economic implications: the credence goods problem is compounded by a cognitive problem in which the cues that normally enable consumers to calibrate their trust are systematically misleading in the AI context.

4.3 The Market for MIRROR Benchmarking

The existence of the MIRROR benchmark itself raises interesting economic questions about the market for AI calibration evaluation. The MIRROR framework represents an attempt to produce public information about AI calibration quality—providing users with standardized metrics for evaluating whether specific AI systems accurately represent their own reliability. From an economic perspective, MIRROR benchmarking is a public good: its results are non-excludable (anyone can use them to evaluate AI systems) and non-rivalrous (one organization's use of MIRROR findings does not reduce their value to others).

The public goods character of MIRROR benchmarking implies that it will be underproduced relative to its social value. The organizations most capable of conducting sophisticated calibration evaluation—large AI developers, research institutions, well-resourced regulators—are not the ones with the strongest incentive to produce benchmarking information that benefits the entire AI ecosystem. This creates a structural role for public investment in AI calibration evaluation

infrastructure: governments, foundations, and international organizations may need to fund the MIRROR-equivalent infrastructure that markets will not produce at optimal levels.

---

5. Integrating MIRROR and Verification Tax: A Structural Analysis

5.1 The Circularity of AI Trust Market Failure

The combination of the MIRROR findings and the Verification Tax analysis reveals a circularity in the AI trust market failure that makes it self-reinforcing. Miscalibrated AI systems produce unreliable outputs that organizations and individuals use to make consequential decisions. The Verification Tax prevents auditing mechanisms from identifying and flagging this miscalibration. Users therefore cannot distinguish calibrated from miscalibrated AI systems, providing no market signal to reward calibration investment. AI developers continue to underinvest in calibration because the market cannot distinguish their calibration efforts. The result is a self-reinforcing equilibrium in which miscalibration persists and the AI trust market remains fundamentally broken.

This circularity is economically significant because it implies that isolated interventions in the AI trust market will be insufficient. Requiring AI developers to publish calibration reports addresses the information asymmetry problem but ignores the Verification Tax constraints that make those reports unreliable. Mandating third-party auditing addresses the credibility problem but ignores the statistical and mechanistic costs that make comprehensive auditing infeasible. Each intervention addresses one dimension of the market failure while leaving the others intact.

More specifically, the HR analytics case illustrates three distinctive economic mechanisms through which AI trust market failure operates. First, the temporal mismatch between AI recommendation and outcome creates a verification lag that prevents efficient market learning. A hiring recommendation based on an AI resume screening may take months or years to reveal its quality through employee performance outcomes—by which time the AI system may have been updated, the evaluator may have left the organization, or the organizational context may have changed. This verification lag prevents the market feedback mechanisms through which ordinary goods markets discipline quality: the delay between purchase and quality revelation means that even the most attentive organizations cannot efficiently learn from experience.

Second, the bundling of AI systems with organizational processes makes it difficult to isolate the AI quality signal from the noise of other organizational variables. An employee's success or failure in a role depends on factors beyond hiring quality—the quality of onboarding, managerial support, team dynamics, and organizational culture all influence performance. This causal opacity means that organizations cannot reliably attribute performance outcomes to AI quality, preventing the market learning mechanism from operating even in principle.

Third, the network effects in AI adoption create a collective action problem that compounds the trust market failure. When an organization adopts an AI system that turns out to be miscalibrated, it faces strong incentives to hide this fact—both because admitting the failure implies reputational costs and because the organizational investment in the AI system creates commitment that resists acknowledgment of failure. This silence prevents the public disclosure of AI quality information that would otherwise enable market learning. The resulting information deficit means that organizations that deploy the same AI system after a predecessor's failure cannot access the failure information that would enable them to make better adoption decisions.

The economic cost of this trust market failure in HR analytics is substantial. Research on the economic value of AI calibration finds that organizations deploying miscalibrated AI systems experience productivity losses that exceed the gains from AI assistance—net negative returns that are not visible in individual organizational calculations because the counterfactual of well-calibrated AI cannot be directly observed. At the macroeconomic level, the aggregation of these organizational-level losses represents a significant drag on human capital allocation efficiency—the misallocation of talent to roles where AI-assisted screening has systematically distorted the selection process.

5.3 The Double Public Goods Problem

A particularly insidious feature of the AI trust market failure is what might be termed the double public goods problem: both the underlying good (calibrated AI) and the remedy for its absence (trust infrastructure) are public goods subject to underinvestment. The calibration of AI systems—making them accurately represent their own knowledge boundaries—is a public good because its benefits are non-excludable (all users of the AI benefit from its calibration) and non-rivalrous (one user's benefit does not reduce another's). The trust infrastructure necessary to verify AI calibration—auditing frameworks, benchmark methodologies like MIRROR, certification institutions—is equally a public good for the same reasons.

This double public goods problem creates a vicious cycle: the underinvestment in calibration produces unreliable AI systems; the underinvestment in trust infrastructure prevents the detection and correction of calibration failures; the resulting AI quality degradation increases the demand for trust infrastructure; but the public goods character of trust infrastructure prevents the market from meeting this demand. Breaking this cycle requires coordinated public intervention that addresses both layers of the public goods problem simultaneously—investing in both calibration standards and the auditing infrastructure necessary to verify compliance with those standards.

The economic literature on coordination failures provides theoretical grounding for why such coordinated intervention is necessary. Multiple equilibria in markets with public goods dynamics mean that the AI trust market can be stuck in a low-trust, low-investment equilibrium that no individual actor can escape unilaterally. Only coordinated collective action—through industry standards bodies, regulatory mandates, or international agreements—can shift the market to a high-trust, high-investment equilibrium in which calibration investment is rewarded and trust infrastructure is adequately funded.

5.2 HR Analytics as a Case Study in Trust Market Failure

The human resource analytics domain examined by Paper 3 provides a concrete case study in how the AI trust market failure manifests in organizational decision-making. Organizations deploying AI for hiring, performance evaluation, and talent management face the full complexity of the credence goods problem: they cannot directly evaluate whether the AI systems they purchase are calibrated for their specific deployment contexts, whether the explanations those systems produce are faithful to actual decision processes, or whether the recommendations they generate are reliable guides to the outcomes they predict.

The economic consequence of this trust market failure in HR analytics is a suboptimal allocation of organizational resources. Organizations that overtrust miscalibrated AI systems make poor hiring and promotion decisions that reduce workforce quality. Organizations that undertrust well-calibrated AI systems forgo the efficiency gains that AI-assisted decision-making could provide. The net effect is a misallocation of human capital—the most valuable resource in modern economies—due to the inability of markets to efficiently price AI reliability.

Research on strategic human resource analytics with explainable AI demonstrates that organizations are aware of the AI trust problem and invest in explainability as a partial remedy. Yet the economic analysis developed in this paper reveals that explainability is itself subject to the same market failure dynamics: organizations cannot efficiently evaluate whether the explanations they receive are accurate, and AI developers have weak incentives to produce genuinely informative rather than merely plausible explanations.

---

6. The AI Trust Infrastructure Framework

6.1 Market Mechanisms for AI Trust

Addressing the AI trust market failure requires a multi-pronged strategy that combines market mechanisms, regulatory standards, and institutional innovations. Several market-based approaches can partially correct the incentive misalignment that drives AI calibration underinvestment.

Performance-based contracting—contracts that tie AI developer compensation to verified AI performance metrics—can align private incentives with social returns on calibration investment. When AI developers bear the costs of miscalibration through contractual liability or performance-based fee structures, their private returns to calibration investment increase. Yet performance-based contracting faces the same measurability challenges that undermine certification markets: if AI quality cannot be feasibly verified, performance-based contracts cannot be written in enforceable terms.

Reputation mechanisms can partially substitute for direct quality evaluation, particularly in repeated-transaction settings. AI developers with strong reputational incentives—those with established brands, long-term customer relationships, and visibility into future transactions—have stronger incentives to maintain calibration quality than those in one-time transaction markets. Yet reputation mechanisms are particularly weak in AI markets because the consequences of AI miscalibration may not manifest until long after the transaction, and because the technical complexity of AI makes reputation based on perceived quality vulnerable to manipulation by sophisticated actors.

6.2 Regulatory Standards and Institutional Infrastructure

The structural nature of the AI trust market failure implies that regulatory intervention is necessary to establish the institutional infrastructure that markets cannot produce. Several regulatory approaches merit consideration.

Mandatory calibration disclosure requirements—requiring AI developers to publish standardized calibration metrics using frameworks like the MIRROR benchmark—address the information asymmetry problem by ensuring that the minimum information necessary for market evaluation is available. The economic challenge is that the quality of disclosed metrics is itself subject to the Verification Tax constraints; organizations cannot efficiently verify that disclosed metrics accurately represent actual AI calibration. Mandatory disclosure must therefore be paired with institutional mechanisms for verifying disclosed metrics.

AI auditing mandates—requiring that high-consequence AI deployments be certified by independent auditors as meeting minimum calibration and reliability standards—address the credibility problem by substituting institutional verification for individual consumer evaluation. The auditing mandate must, however, grapple with the Verification Tax constraints that limit the comprehensiveness of any individual audit. This implies that auditing mandates should focus on

the most consequential AI applications—the domains where the social cost of miscalibration is greatest—rather than attempting comprehensive coverage of all AI deployments.

International coordination is essential for the effectiveness of AI trust regulation. AI systems operate across national borders, and regulatoryFragmentation—different standards in different jurisdictions—creates both compliance costs and regulatory arbitrage opportunities that undermine the effectiveness of national-level interventions. International coordination mechanisms, such as mutual recognition agreements for AI auditing standards and shared frameworks for AI calibration measurement, can reduce the coordination costs of effective AI trust governance.

6.3 Institutional Innovations for Trust Infrastructure

Beyond market mechanisms and regulation, institutional innovations can address the structural barriers to AI trust. Several promising approaches merit attention.

Calibration certification markets—specialized intermediaries that focus exclusively on evaluating and certifying AI calibration quality—can address the adverse selection problem by concentrating expertise and creating standardized metrics that enable comparison across AI systems. Certified calibration quality becomes a market differentiator that rewards calibration investment. The economic challenge is that certification quality is itself subject to the Verification Tax constraints; certification intermediaries must invest in the same statistical and mechanistic evaluation capacities that make comprehensive auditing expensive.

AI calibration research funding—public investment in calibration methodology development—addresses the public goods problem in calibration infrastructure by ensuring that the foundational tools for calibration measurement are produced at socially optimal levels. Research investment in MIRROR-equivalent benchmarking frameworks, calibration training techniques, and calibration evaluation methodologies benefits the entire AI ecosystem but will not be produced by private actors at optimal levels.

Trust arbitration mechanisms—institutional processes for adjudicating disputes about AI quality—can address the liability gap by creating credible enforcement mechanisms that make contractual and regulatory commitments on AI quality enforceable. When organizations can hold AI developers accountable for documented miscalibration harms, the incentive to invest in calibration increases.

---

7. Conclusion

This paper has argued that the AI trust challenge is fundamentally an economic problem arising from the interaction of three market failures: the credence goods problem that prevents consumers from evaluating AI quality, the public goods problem that leads to systematic underinvestment in AI calibration, and the verification market failure that prevents efficient certification of AI reliability. The three foundational references illuminate different dimensions of this economic phenomenon. The MIRROR benchmark (Paper 1) documents the output of the calibration underinvestment problem: state-of-the-art AI systems are systematically miscalibrated because the incentive structures they face do not reward calibration investment commensurate with its social value. The Verification Tax analysis (Paper 2) reveals the structural constraints that make the verification market failure inherent rather than incidental: the resource costs of comprehensive AI evaluation cannot be reduced below certain floors by methodological improvement alone. The strategic HR analytics framework (Paper 3) demonstrates how these abstract economic dynamics manifest concretely in organizational decision contexts where the AI trust market failure imposes real costs on organizations and individuals.

The practical implication is that addressing the AI trust challenge requires more than technical solutions to AI reliability; it requires institutional innovations that restructure the economic incentives surrounding AI development, deployment, and governance. Markets alone will not produce calibrated AI trust because the structural conditions for efficient market operation—measurability, excludability, and adequate incentive alignment—are systematically absent in AI markets. The AI Trust Infrastructure Framework proposed in this paper provides a roadmap for the coordinated institutional intervention that the AI trust market failure demands: market mechanisms to align private incentives with social returns on calibration, regulatory standards to establish minimum quality floors and ensure information availability, and institutional innovations to address the verification and certification challenges that market mechanisms alone cannot solve.

References

1. Einav, L., & Knoeper, J. D. (2025). Credits and credence: A paradigm for credence goods markets. American Economic Review.
2. Vanneste, B., & Puranam, P. (2024). Artificial intelligence, trust, and perceptions of agency. Academy of Management Review.
3. Wang, J. Z. (2026). MIRROR: A Hierarchical Benchmark for Metacognitive Calibration in Large Language Models. arXiv preprint arXiv:2604.19809.
4. Emons, W. (2024). Credence goods: A survey of the empirical literature. Journal of Economic Surveys, 38(1), 152-179.
5. Arrow, K. J. (1972). Gifts and exchanges. Philosophy & Public Affairs, 1(4), 343-362.
6. Wang, J. Z. (2026). The Verification Tax: Fundamental Limits of AI Auditing in the Rare-Error Regime. arXiv preprint arXiv:2604.12951.
7. Bei, J., Liu, Z., Huang, J., Wang, X., & Yang, P. (2025). Strategic Human Resource Analytics with Explainable Artificial Intelligence. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology.
8. Glikson, E., & Woolley, A. W. (2025). Organizational trust in AI: A review and research agenda. Academy of Management Annals.
9. Lanzolla, G., et al. (2024). The rise of AI: Issues for theory and practice. Academy of Management Annals.
10. Klingbeil, J., et al. (2024). Trust formation, error impact, and repair in human–AI financial advisory: A dynamic behavioral analysis. PMC, National Institutes of Health.
11. Diethe, T., et al. (2024). Quantifying the economic value of AI calibration. Nature Machine Intelligence.
12. Müller, V. C., & Bostrom, N. (2024). Fundamental issues of artificial intelligence: A survey of relevant frameworks. Synthese.
13. Bernstein, A., & Mnookin, J. (2024). Market design for AI trust. Harvard Journal of Law and Technology.
14. Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In Uncertainty in Economics (pp. 235-251). Academic Press.
15. European Commission. (2024). The EU Artificial Intelligence Act: Economic and social implications of trustworthy AI. Official Publications of the European Union.