Scientific Research Publishing

American Journal of Industrial and Business Management, 2026, 16(8), 826-847
https://www.scirp.org/journal/ajibm
ISSN Online: 2164-5175
ISSN Print: 2164-5167

Governing the Machine: AI Ethical Frameworks and Governance Boards

Sidaarth Karegowdra

Independent Research GT, River Hill High School, Clarksville, USA
Email: karegowdrasidaarth@gmail.com

How to cite this paper: Karegowdra, S. (2026). Governing the Machine: AI Ethical Frameworks and Governance Boards. *American Journal of Industrial and Business Management, 16,* 826-847.
https://doi.org/10.4236/ajibm.2026.168044

Received: July 13, 2026
Accepted: August 2, 2026
Published: August 5, 2026

Abstract

Artificial intelligence is being incorporated into high-stakes sectors of society such as healthcare, employment, and criminal justice at a rate that outpaces governance structures designed to mitigate the potential risks that come with the novel technology. To that end, despite its widespread usage, only 31% of organizations globally maintain formal AI governance policies, which has paved the way for a multitude of cases on bias concerns, privacy violations, and decreasing public trust. This study explores whether a mandatory, government-centered AI governance framework with Explainable AI (XAI) requirements is associated with stronger public trust outcomes than voluntary, industry-led approaches. Additionally, it aims to identify whether independent governance boards with genuine enforcement authority within organizational settings are associated with improved accountability. A structured narrative synthesis of six peer-reviewed sources was conducted in an effort to compare governance mechanisms and oversight models by drawing on survey-based, experimental, organizational, and theoretical samples. In order to evaluate the results of the studies across different sectors, effect sizes were documented and compared descriptively in their original reported metric, rather than converted to a common metric and statistically pooled. Two summary tables translated these data points into a structured comparison across sources. Across the six sources, XAI/transparency conditions were consistently associated with a small to moderate positive relationship with public trust, while a coercive, compliance-based mechanism examined in one organizational study was associated with the weakest outcome of any mechanism reviewed. Trust transfer dynamics further supported that human intermediaries can meaningfully drive AI trust given that expertise and credibility are maintained. Ultimately, these findings suggest that while Explainable AI requirements are associated with a small to moderate positive relationship with public trust, the maintenance and long-term strength of that trust may depend more on the presence of enforceable governance structures than on transparency alone. Because none of the six

sources directly compares mandatory government-centered governance to voluntary industry-led governance, or directly tests the effect of independent board authority, these findings should be interpreted as associative evidence rather than as a direct causal test. These results tie together to inform policy recommendations aimed at proposing Maryland legislation encouraging XAI standards and independent AI governance boards, and also guiding outreach to small AI companies.

Keywords

Artificial Intelligence (AI), AI Governance, Explainable Artificial Intelligence (XAI), Government Regulation, Transparency, Narrative Synthesis

1. Literature Review

1.1. Introduction

Artificial intelligence now influences decisions that directly affect people's lives, including hiring decisions, medical diagnoses, credit approvals, and criminal sentencing. Not surprisingly, organizations have begun to use AI in institutional tasks such as sorting job applications and analyzing lists of data to predict consumer trends. However, despite AI's rapid growth in society, sufficient safeguards and AI governance models have not been adequately developed to match the transition. Reports indicate that the majority of organizations around the globe are deploying some form of AI, yet only 31% have formal policies guiding ethical development and implementation (ISACA, 2025). The United States currently relies on voluntary framework models to regulate AI usage across various industries, sparking inconsistency in compliance and increased risk of bias, discrimination, and poor decision making. Specifically, the U.S. has been using mechanisms such as non-mandatory audits, vague government policies, and internal review processes that lack formal authority. Corporations cannot be fully entrusted to monitor their AI use because ulterior motives such as profit, market competition, and innovation may take priority over public responsibility. Centralized enforcement and structural reform offers a pathway to avoid a future with public trust in jeopardy, and helps convert written policy into efficient daily practices for public entities to follow. A universal government-centered AI ethical framework, paired with independent AI governance boards within U.S. public corporations, may strengthen transparency and accountability in algorithmic systems by establishing enforceable oversight, standardized auditing practices, and consistent public-facing reporting, thereby improving public trust in corporate AI governance. The purpose of this paper is to analyze international models of AI governance, outline the essential components of an effective government-centered universal AI ethical framework, and evaluate the specific structure and responsibilities of corporate AI review boards in order to show how these combined mechanisms may help ensure fair and transparent AI practices in U.S. public corporations.

1.2. Background

In recent years, the use of artificial intelligence has skyrocketed in public-facing and corporate sectors, leading to both opportunity and ethical concerns. Specifically, artificial intelligence has become a significant component in industries such as healthcare, finance, and education, due to its ability to provide efficiency, automation, and predictive decision-making. Students can use it as a source of assistance on challenging tasks, investors can use the tool to analyze market trends, and it is increasingly evolving in surgical procedures, even replacing humans in some instances. However, this expansion brings about ethical risks such as bias, privacy violations, and lack of accountability (Hadley et al., 2024). In order to theorize foundational principles or methods to mitigate potential harm, technical experts and ethicists turn to early frameworks built around the concept of ethical research, such as the Belmont Report, a foundational document in American history that outlines key principles and guidelines for protecting human subjects in research. These frameworks often highlight aspects of fairness like respect for persons, beneficence, and justice. Aspects of fairness laid out from decades ago continue to influence ethics through modern approaches in AI, despite the fact that technology was drastically different between time periods.

Throughout the duration of this paper, it is important to characterize key terms that are instrumental in gauging the understanding of frameworks and review boards. Artificial intelligence is a field of computer science that centers around creating machines capable of performing tasks that would normally require human intelligence, including but not limited to decision making, language detection, problem solving, and social interaction (Schuett et al., 2024). Algorithmic Review Boards (ARBs) are governance bodies within corporations that are tasked with implementing and executing ethical AI principles into practice, monitoring compliance with regulations that are upheld, managing risks, implementing whistleblower channels, applying bias mitigation strategies, and more (Hadley et al., 2024). Lastly, Algorithmic Impact Assessments (AIAs) are well developed and structured tools that are designed to catch any risks or harms that are emerging from recently developed automated decision-making systems and assess risk of discrimination or privacy breaches (Treasury Board of Canada Secretariat, 2021).

There are several mechanisms that have been proposed to improve ethical AI governance; however, inconsistencies in these approaches arise in the face of reliance on voluntary frameworks. If the frameworks that are supposed to uphold these crucial mechanisms are optional to follow, corporations have very little incentive to actually put them into practice due to profit-driven mindsets. In addition to voluntary frameworks, algorithmic review boards with limited decision-making authority also contribute to why mechanisms are inconsistent. Both of these pitfalls lead to risk of a lack of transparency in areas such as hiring, credit, and criminal justice, potentially costing individuals massive amounts of income or job opportunities.

Clearly, public concern about AI ethics and transparency continues to grow as

a result of day-to-day life being severely impacted by negative repercussions of unethical systems. In order to regain public trust and address system discrimination, this paper argues that implementing universal AI ethical frameworks combined with independent corporate governance boards is a promising step, while acknowledging that the empirical evidence synthesized in this paper is associative rather than a direct experimental test of that claim (see Methods).

1.3. Role of International AI Governance Frameworks in Trustworthy Systems

International AI governance frameworks show that risk-based regulation, mandatory audits, and enforceable standards are associated with more trustworthy and accountable systems. It is imperative to analyze non-U.S. and global models in order to gain a better understanding of effective AI governance frameworks and identify common mistakes to avoid. Canada's Directive on Automated Decision-Making requires all federal agencies deploying automated decisions to complete an Algorithmic Impact Assessment prior to use. Each system is scored using levels, with level I indicating little to no impact and level IV indicating extremely high impact. This requirement is an obligation by the nation and ensures that no automated system can be used by the public without prior in-depth analysis and approval. Canada is able to avoid the regulatory failure of selective or voluntary compliance by applying universal oversight as well as scaling requirements by risk level. Level III and level IV systems on the risk scale draw the most scrutiny from critics and technical experts, and are present in highly impactful sectors such as healthcare and banking. In order to reduce the chance of high-level impact, Canadian agencies must publish public-facing explanations of system logic, conduct external review, retain human-in-the-loop decision-making authority (humans are the ultimate decision makers, not AI), and perform continuous post-deployment monitoring (Choung et al., 2023). A clear trend emerges with the growing number of obligations and regulations that highlights how mandatory audits can elevate "textbook" practices to genuine, measurable actions that implement transparency and oversight. Audits shape explanations in a way that is easier to comprehend for public viewing, which reduces confusion between government institutions and people affected by AI systems. External review and human-in-the-loop decision-making authority builds credibility of deployed AI systems because it helps ensure that AI is never the sole decision maker in consequential decisions or goes unchecked in the deployment stage itself. These measures are further reinforced through continuous post-deployment monitoring, which helps ensure that even after AI systems are put into practice, experts keep track of their performance.

Similar to Canada's mandatory auditing approach, one of the most pivotal pieces of work on AI ethics is the European Union's "trustworthy AI" framework, which shows how ethical principles can be translated into concrete regulatory obligations. The EU High-Level Expert Group on AI outlines ethical principles across seven dimensions of "trustworthy AI", including human agency

and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal well-being, and accountability (Madiega, 2019). The initial stage of developing AI systems is instrumental in that it lays out criteria for certain ideals and how to evaluate them in practice. Notably, the identification of specific dimensions that can be evaluated during system development and deployment, instead of framing AI as a list of values (as seen in earlier framework models like the Belmont Report), sets a precedent for future frameworks on how trustworthy systems are more likely to emerge when ethical principles are built into governance structures instead of being left for broader interpretation by corporations. Additionally, the framework unites the diverse groups involved in the development and execution of frameworks, such as developers, regulators, and the public, by creating shared expectations of what "trustworthy AI" entails while decreasing ambiguity. Without this type of unity and formality, ethical guidelines are insufficient, which is clearly highlighted in an Information Systems Audit and Control Association (ISACA) survey that found that while the majority of IT and business professionals in Europe believe employees in their organization are using AI, only 31% of those organizations have a formal, comprehensive AI policy in place (ISACA, 2025). Ethical guidelines alone are not enough. Developers may place insufficient emphasis on transparency and oversight because of factors like cost, complexity, and the potential for a loss of focus on obtaining profit. Voluntary compliance is limited, so the EU takes steps to fill the gap of optional ideals by backing up trustworthiness with enforceable obligations.

The risk-based classifications under the proposed EU Artificial Intelligence Act further display how regulation can be proportionate while still comprehensive. This proposed act is similar to the aforementioned Algorithmic Impact Assessment obligations from Canada's Directive, as it categorizes AI systems into four risk tiers (minimal, limited, high, and unacceptable risk), with high-risk systems accounting for a minority of AI applications yet representing the majority of potential legal and societal harms (Ruschemeier, 2023). This classification system tends to focus more on impact than on technical development so that the EU can avoid overregulating harmless applications while directing more attention toward systems with greater involvement in employment and civil liberties. These sectors require greater security and targeted oversight, as they are classified as high-risk systems. A common criticism of AI regulation is that it limits innovation, but these classifications take steps to focus heightened accountability specifically where it matters most. However, classification should not be mistaken for leniency, as high-risk systems must undergo pre-market conformity assessments, maintain detailed technical documentation for extended periods, and comply with post-market monitoring requirements across all 27 EU member states (Veale & Borgesius, 2021). The EU's requirements standardize oversight across a large multinational audience in order to help prevent regulatory arbitrage, in which companies take advantage of weaker enforcement elsewhere to increase profit. Long-term documentation obligations improve audita-

bility by allowing regulators to reconstruct system behavior years after deployment, while pre-market assessments help identify risks before systems become widely used in consequential decisions.

Independent algorithmic auditing research demonstrates that mandatory external oversight is important for surfacing harms that self-regulation fails to catch. The Gender Shades project illustrates the value of audits in detecting bias. The project team audited three commercial facial recognition systems produced by IBM, Microsoft, and Face++, and found substantial disparities in error rates across gender and skin color, citing an error rate of 34.7% for dark-skinned women (Ugwudike, 2021). As a result of this audit, the companies involved took measurable steps to address bias in their systems. Algorithms that influence public life "deserve scrutiny", as audit studies reveal that discriminatory outcomes can persist even when systems are presented as objective (Sandvig et al., 2014). A significant limitation of corporate self-regulation lies in the fact that companies have relatively little incentive to uncover bias in proprietary systems on their own, as they tend to place a higher emphasis on production than on oversight, and may not make adequate efforts to address bias until legal or reputational consequences arise. The findings further support the idea that trustworthy AI systems benefit from external accountability, either through peer review or human-in-the-loop decision making, that can operate independently of corporate interest in profit or sales. Surveys and qualitative reviews of corporate AI practices, featuring industry practitioners, policy leads, and members of internal ethics teams, suggest that only a small minority of corporate AI ethics initiatives include independent audits, while the majority rely on internal ethics committees without enforcement powers (Bietti, 2019; Raji et al., 2020). Lack of enforcement and external review are cited as contributing factors in cases where voluntary governance has resulted in weaker protections across sectors. For example, Amazon's AI hiring tool demonstrated discriminatory patterns against female applicants because it downgraded resumes containing terms associated with women, illustrating how internal systems can still produce biased outcomes even without external pressure to change (Langenkamp et al., 2019; Dastin, 2018). This suggests that internal ethics committees may benefit from additional external authority or transparency mechanisms to more effectively counterbalance profit-driven decision making.

International frameworks demonstrate the potential value of structured governance backed by strong enforcement mechanisms and regulatory obligations, while also pointing to apparent limitations of voluntary commitments through the various studies discussed above on mandatory audits and enforceable standards. Building upon this history of international frameworks offers one path for the United States to develop a domestic equivalent, which motivates the outline of an efficient government framework discussed below.

1.4. Debates Surrounding Mandatory AI Regulation

Although mandatory oversight and enforcement mechanisms have shown prom-

ise in mitigating harms within algorithmic systems, some may argue that centralized enforcement risks slowing innovation. The Computer and Communications Industry Association (CCIA Europe) has warned that premature regulatory enforcement can be detrimental to innovation and consumer choice in generative AI (Davies, 2024). The CCIA, in reference to the EU AI Act, has expressed concern that AI developers may be overburdened with disproportionate compliance costs. If harsher regulation is in place, partnerships between big-tech companies and smaller start-ups could increase as a result of insufficient funding for smaller firms to keep up with regulatory costs, and critics worry that larger companies could use such partnerships to gain outsized influence over smaller firms' technology. Proponents of this view suggest that exploitation through partnerships and reduction in consumer choice could be reduced if regulation like the EU AI Act is implemented more gradually. However, innovation is not always hindered by regulation, as historical evidence from environmental and data protection law suggests that regulation can also drive new forms of privacy-protective and eco-friendly technology. Additionally, legal uncertainty itself can decrease investment confidence and contribute to market fragmentation. Standards developed by the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) can help ensure consistency and support innovation by enabling scalability and lowering market uncertainty. Therefore, the notion that regulation is a purely static restraint on growth is contested; it can also function as a structural mechanism that supports more sustainable, predictable growth.

1.5. An Outline of an Efficient Government Framework

An effective United States AI governance framework would likely benefit from measurable, concrete, enforceable standards rather than voluntary guidelines alone, in order to foster more consistent accountability. The Federal AI Governance and Transparency Act is a proposed legislative effort introduced in the House of Representatives on March 5, 2024, that identifies transparency, risk assessment, and institutional oversight as major elements of federal AI governance. The legislation would require agencies to document both the deployment and development of automated systems used in decision-making processes (H.R. 7532, 2024). This approach is intended to ensure that transparency functions as a mechanism for accountability rather than as an aspirational, unenforced commitment. These requirements would operate at the federal level, which proponents argue would reduce agency discretion over whether to disclose AI use and support more consistent institutional self-interest management. The legislation also proposes creating a Federal AI System Inventory, which would require agencies to document AI systems currently in use, their intended purposes, and the decision contexts in which they operate (H.R. 7532, 2024). Such an inventory could serve as a centralized oversight mechanism, allowing regulators and the public to identify high-risk areas of AI deployment and helping policymakers direct regulatory effort toward

systems more likely to produce harmful outcomes. The act also would require agencies to assign human oversight authorities and establish appeal mechanisms for individuals affected by algorithmic decisions, intended to help ensure that automated systems remain subject to human accountability.

Building on the case for regulating accountability through frameworks, mandatory audits and impact assessments are proposed as mechanisms to help expose harms that voluntary compliance may fail to catch. Proponents of the Algorithmic Accountability Act argue that mandatory impact assessments can help reveal discriminatory, confusing, or privacy-harming systems that voluntary approaches often fail to identify (Mithal, 2022). A central concern in risk analysis is avoiding an emphasis on protecting institutional reputation over conducting in-depth risk review; voluntary frameworks are argued to be more prone to this tendency, which proponents suggest can contribute to weaker long-term oversight. Related research indicates that systems required to undergo mandatory audits are more likely to document training data sources, decision logic, and mitigation strategies compared to systems governed by internal ethics policies alone (Morley et al., 2021). Audits are one form of documentation that can support external evaluation and traceable accountability, whereas organizations relying solely on internal ethics policies may have less standardized, less traceable verification processes. Organizational studies find that relatively few institutions relying on voluntary AI ethics frameworks conduct third-party audits, while mandatory approaches typically require review at pre-defined stages of development and deployment (Schuett et al., 2024).

Mitigating common enforcement loopholes and reinforcing the potential benefits of mandatory practice would likely require federal oversight to help ensure consistent compliance across agencies. The Government Accountability Office has documented cases where poorly governed federal AI systems produced inconsistent or harmful outcomes in areas such as automated hiring and benefits eligibility determinations (Accountability Framework, 2021). The GAO does not identify the specific AI systems or agencies involved in each case but highlights a broader pattern of risks arising from bias, data quality issues, and challenges complying with federal policies, illustrating how inconsistent oversight can weaken governance structures. One proposed response to these gaps is Inspector General oversight, intended to channel responsibility more clearly through independent enforcement authority (Accountability Framework, 2021), on the reasoning that third-party audits may offer more consistently transparent reporting than internal review alone. Similarly, the NIST AI Risk Management Framework offers guidance on efficient governance practices but remains nonbinding and does not include penalties for non-adoption (AI RMF Development, 2021), which some argue limits its practical effect on organizations that might otherwise selectively follow its recommendations.

Taken together, federal AI governance investigations, studies, and analyses suggest that mandatory enforcement mechanisms and centralized oversight may be

important for translating ambitious governance goals into measurable requirements for government agencies to follow. Internal review boards, discussed next, are one mechanism proposed to help translate written guidelines into daily compliance practice by absorbing new knowledge about AI policy and adjusting internal procedures over time.

1.6. Evaluating the Effectiveness of Corporate AI Review Boards

Corporate-level AI review boards are one mechanism proposed to translate abstract ethical values into enforceable, day-to-day decision-making processes. AI review boards can provide an internal governance mechanism that connects organizational practice with external legal and ethical standards. Corporate review boards are interdisciplinary entities within organizations that monitor the development, use, and long-term impact of algorithmic systems. One prominent form is the Algorithmic Review Board (ARB), which draws on experts from diverse backgrounds to help ensure functional and legal compliance from pre-deployment through post-deployment monitoring. Although review boards are sometimes perceived as complex or burdensome to implement, interviews with ARB practitioners suggest that they may operate most effectively when integrated into existing compliance and risk management structures rather than functioning as standalone ethics committees (Hadley et al., 2024). Embedding ethical oversight within existing organizational processes may encourage institutions to evaluate legal, operational, and ethical risks together rather than separately, helping ethical review become part of routine decision-making rather than an optional add-on step. ARBs are commonly implemented in highly regulated sectors such as finance, healthcare, and insurance, where structured oversight mechanisms already exist (Hadley et al., 2024). For example, JPMorgan utilizes Model Risk Management (MRM) frameworks in which committees regularly review risk models before deployment (Edgerton & Oktem, 2025; Wallen, 2025). MRM frameworks function similarly to ARBs by providing oversight, requiring documentation, and approving models based on formal risk assessments. A comparable emphasis on interdisciplinary governance can be observed in healthcare institutions such as the Mayo Clinic, which relies on multidisciplinary teams, including governance committees, information technology specialists, and legal and executive leadership, to oversee regulatory compliance, technological implementation, and ethical standards (Loufek et al., 2024). The presence of diverse expertise may strengthen institutional credibility and improve the ability of review boards to translate complex technical concepts into clear, publicly understandable guidance.

Clearly defined structure, membership, and authority appear to be key factors in whether AI review boards can function as genuine enforceable governance bodies rather than as a symbolic checkmark for ethical initiatives. Schuett et al. (2024) identify five core design dimensions of effective governance boards: responsibilities, legal structure, membership, decision-making authority, and resources. Ethical oversight is unlikely to be effective when boards lack formal authority within

organizations, and ethical principles risk becoming forgotten commitments absent designated decision-making power. Explicit mandates that clarify who evaluates risk, who makes final determinations, and how decisions are executed are one proposed solution to this problem. Schuett et al. (2024) further emphasize the importance of diverse professional backgrounds and clearly defined decision-making procedures for evaluating bias and risk across different systems and teams. Multidisciplinary membership can help ensure that ethical review extends beyond technical performance to also consider customer impact, and formal governance models may reduce structural ambiguity within an organization. When these two elements, formal authority and diverse membership, are combined, review boards are more likely to function as stable governance entities rather than as discussion forums alone.

Because AI review boards are internalized within organizations, organizational culture and internal accountability mechanisms are also important indicators of daily ethical decision making. Gambelin (2024) writes that "you can't have responsible AI without ethics, and you can't have ethics without values", underscoring that organizational culture must be embedded in both people and technology in order to meaningfully influence behavior. This perspective suggests that review boards matter beyond their written mandates: ethical principles are likely to have a more lasting impact when they are built into organizational roles and everyday procedures rather than treated as periodic compliance checks. Documented corporate failures, such as Amazon's AI hiring tool, which was abandoned after internal testing revealed bias against women, are frequently cited as examples of the consequences of insufficient pre-release review and limited internal escalation channels (Dastin, 2018). Cases like this illustrate how ethical failures can occur even in the absence of malicious intent, when organizations lack functioning oversight and accountability mechanisms capable of identifying early signs of bias.

Overall, this section frames corporate AI review boards as one way of treating responsible AI as a governance problem rather than a purely technical one. If review boards are given formal structure, clear authority, and interdisciplinary participation, the literature reviewed here suggests that ethical principles are more likely to be meaningfully applied in daily decision making, though the empirical support for this claim, discussed further in the Methods and Results sections, comes primarily from organizational case studies and design frameworks rather than controlled comparisons of board authority.

1.7. Conclusion

The current issues in AI governance reflect the limitations of existing enforcement mechanisms. As AI works its way into healthcare, employment, education, and banking, the literature reviewed above suggests that voluntary frameworks alone may not offer sufficient enforcement to maintain a high standard of compliance. When compliance is optional, corporations may weigh profit and market competition more heavily than public responsibility. Concrete, centrally enforced ethical

guidelines are proposed in this literature as a way to support both the generation and the long-term maintenance of accountability. A government-centered framework combined with independent corporate Algorithmic Review Boards is presented in this paper as one path toward more consistent oversight and clearer responsibility, drawing on international models that illustrate mechanisms such as risk-based regulation and mandatory audits. The empirical synthesis that follows in the Methods and Results sections tests a narrower, more specific piece of this argument: whether transparency and institutional credibility are associated with public trust in AI systems, and what that association may imply for governance design going forward.

2. Methods

This study investigates the research question "To what extent does a mandatory, government-centered AI governance framework incorporating Explainable AI (XAI) requirements produce stronger public trust outcomes than voluntary, industry-led approaches, and how does the inclusion of an independent governance board with enforcement authority further improve accountability?" None of the six sources synthesized below directly manipulates governance structure as an independent variable within a single study (for example, by comparing a mandatory, board-certified condition against a voluntary condition head-to-head), so this question is addressed here through converging, associative evidence rather than a single direct statistical test. The hypothesis designed to address this research question focuses on two main aspects. The first aspect delves into requirements pertaining to transparency, focusing on XAI and the link to stronger and more reliable trust outcomes compared to voluntary self-governance. The second aspect centers around the association between independent governance boards with real authority within organizations and improvement in accountability, by transferring the institutional credibility of the oversight body to the AI systems being governed. The combination of third-party governance and clear documentation requiring XAI aims to address the trust deficit that is currently prevalent regarding AI in the United States.

The hypotheses were operationalized into measurable theories by defining "public trust" as reported trust scores or behavioral trust indicators across each included study, and "governance strength" as the level of enforcement present in each institutional framework examined. Both aspects of the hypothesis were addressed through a structured narrative synthesis of six peer-reviewed sources. All of these sources tie together on the core variables of public trust in AI, transparency through explainability, and enforcement/governance structure. Quantitative effect sizes were extracted from the data and reported in their original metric (Pearson r or standardized β, as reported by each source). Because the six sources differ substantially in design, population, and measurement instrument, these effect sizes were compared descriptively across studies rather than converted to a common metric and statistically pooled into a single combined estimate. Two

summary tables were developed from the extracted data to illustrate cross-study comparisons (see Results).

Candidate sources were identified through Google Scholar using the search terms "AI governance", "explainable AI trust", "algorithmic accountability", "AI ethics board", and "ethical AI legislation". An initial pool of approximately 35 peer-reviewed sources was screened, and study selection followed predefined inclusion criteria requiring (1) empirical measurement of trust or governance outcomes, (2) explicit or measurable XAI or transparency variables, and (3) enough quantitative data to compute or extract standardized effect sizes. Six sources meeting all three criteria were retained for synthesis. Effect sizes and supporting data were extracted by the student researcher and checked for accuracy by a faculty advisor (Ms. Sharbaugh). When a source reported multiple outcomes, the outcome treated as primary was identified by determining what the paper was fundamentally trying to show, defaulting to the first or main result when several were reported, and setting aside secondary or exploratory analyses; when authors did not specify a primary outcome, the outcome emphasized in the abstract, or the outcome with the longest follow-up period, was used instead. Because Cheung and Ho (2025) and Li et al. (2021) each report multiple outcomes drawn from the same underlying sample, these related outcomes are presented together as within-study comparisons in the Results section, rather than as independent data points, so that correlated estimates from a single sample do not receive disproportionate weight.

Two of the six sources provided survey-based evidence on public attitudes and XAI effects. Bullock et al. (2025) conducted a nationally representative U.S. survey that examined regulatory preferences. Some statistical evidence pulled from the survey reported standardized beta coefficients (ranging from 0.49 to 0.59) for perceived risk as the strongest predictor of regulatory support. Cheung and Ho (2025) focused primarily on a specific sector of engineering and the use of AI within niche environments. The study's team utilized a cross-sectional public opinion survey of 1,002 stratified respondents in Singapore. They measured the results of the survey by using Structural Equation Modelling (SEM) through MPlus 8.3 to measure how perceived AI explainability relates to trust in AI engineers across ability, benevolence, and integrity.

Two sources provided experimental evidence on XAI and trust calibration. Hao et al. (2026) used a 2 × 2 between-subjects design crossing AI explainability with interaction outcome (success vs. failure) on a population of 120 university participants, alongside Grad-CAM-based explanations and Echo State Networks (ESNs) to model implicit trust dynamically. Ahn et al. (2021) used two controlled experiments to test whether interpretability or outcome feedback is more effective at boosting behavioral trust. A fifth source, Li et al. (2021), employed a cross-sectional field survey of 180 engineers at a petrochemical enterprise in eastern China that had recently deployed an AI fault detection system, examining management commitment, directive leadership, and AI promoter trust as predictors of trust

within AI systems.

The last source, Li et al. (2024), is a systematic review that combines research on interpersonal and human-automation trust across different sectors of society (healthcare, business/finance, civil services) and AI domains. It proposed a three-dimension governance framework that covers measurable characteristics of trust (the trustor), trustee design standards, and contextual legal accountability. Most notably, it is the only source among the six sources that discusses the need for a third-party independent oversight body to audit AI companies. Synthesizing these sources together, this synthesis covers data from survey, experimental, organizational, and theoretical methodologies, which permitted a structured, theme-based comparison of effect sizes across the six sources.

3. Results

The data collected across all six sources demonstrates consistent patterns across both governance structure and public trust outcomes. The evidence cited in Bullock et al. (2025) is consistent with the first component of the hypothesis, regarding mandatory government-centered frameworks and transparency-related outcomes as compared to voluntary approaches, though as a correlational survey it reflects existing public attitudes rather than a direct test of mandatory versus voluntary governance. Bullock's nationally representative U.S. sample found that 42% of Americans lack basic trust in AI, and that trust in government (not trust in AI companies) is the primary predictor of regulatory support. This finding suggests lower relative support for voluntary industry self-governance and aligns with a government-centered framework as the preferred oversight approach for AI in the American public's view. Li et al. (2024) build off this narrative and further support this stance by discussing the need for a neutral third-party oversight authority with genuine enforcement authority within organizations to ensure that proper and frequent audits are conducted, documenting each stage of the process from development to deployment.

The second aspect, focusing on the link between accountability and independent governance boards through trust transfer, is directly supported and touched upon by Li et al.'s (2021) China Manufacturing Study. The dominant predictor of employee trust in AI was trust in the AI promoter ($\beta = 0.532$, $p < 0.001$). An AI promoter is a human intermediary who developed an AI system within the organization that currently uses its models. These findings indicate that credibility associated with human intermediaries is positively correlated with trust in AI systems. Logically, if clients and users do not have knowledge on AI systems they tend to shift their attention and thus rely on the credibility of the institution overseeing those systems. Because management commitment ($\beta = 0.192$, $p < 0.01$) and directive leadership ($\beta = 0.129$, $p < 0.05$) are drawn from this same 180-engineer sample, they are presented alongside AI promoter trust as related, within-study predictors (**Table 1** below) rather than as independent effect sizes.

Table 2 summarizes the reported associations between XAI/transparency and

trust in AI engineers from Cheung and Ho's (2025) Singapore survey. Because all three outcomes are drawn from the same 1002-respondent sample and structural equation model, they are presented together below as a within-study comparison rather than as three independent effect sizes.

Table 1. Reported institutional predictors of trust in AI (Li et al., 2021; N = 180, single sample).

Predictor (Mechanism Type)	Standardized β	Significance	Mechanism Category
AI Promoter Trust	0.532	$p < 0.001$	Institutional/human-intermediary credibility
Management Commitment	0.192	$p < 0.01$	Institutional/voluntary
Directive leadership	0.129	$p < 0.05$	Coercive/compliance-based

Table 2. Reported associations between XAI/Transparency and trust in AI engineers (Cheung & Ho, 2025; N = 1002, single sample).

Outcome (Trust Dimension)	Standardized β	Significance	Design
XAI → Integrity	0.28	$p < 0.001$	Cross-sectional survey, SEM
XAI → Ability	0.15	$p < 0.01$	Cross-sectional survey, SEM
XAI → Benevolence	0.13	$p < 0.001$	Cross-sectional survey, SEM

All three pathways in Table 2 are positive and statistically significant, with the strongest association appearing for the integrity dimension of trust. Notably, of the three trust dimensions, Cheung and Ho (2025) found that only trust in ability significantly predicted downstream attitude toward AI-based systems, which the authors describe as a competence bias in public trust formation: the public may weight perceived technical competence more heavily than perceived integrity or benevolence when deciding whether to rely on an AI system, even though explainability's strongest relationship is with the integrity dimension.

Table 1 summarizes the three within-study associations from Li et al.'s (2021) China Manufacturing sample discussed above, organized by the type of mechanism each represents (institutional/human-intermediary credibility, institutional/voluntary, and coercive/compliance-based).

Read together, Table 1 and Table 2 describe six associations across two independent samples (Singapore and China), all positive, with the coercive/compliance-based mechanism, directive leadership, sitting at the floor of the distribution and producing the smallest reported association with trust of any outcome examined in this synthesis. This distribution suggests that both human intermediary factors and transparency-related variables contribute to trust outcomes across studies, and that coercive enforcement is comparatively the least effective

mechanism for building genuine trust. Because these six values are drawn from only two underlying samples and use two different standardized metrics, they are presented here as descriptive comparisons rather than pooled into a combined statistic.

Hao et al. (2026) provide additional experimental evidence on XAI and trust calibration using a 2 × 2 between-subjects design crossing AI explainability with interaction outcome. In a comparison of the CIFAR-10 image classification task with and without Grad-CAM-based explanations, participants reported significantly higher trust when explanations were present ($M = 5.92$, $SD = 0.43$) than when they were absent ($M = 5.12$, $SD = 0.51$). Separately, Hao et al. also report that explicit (self-reported) and implicit (behavior-based) trust measures were strongly correlated with one another ($r = 0.78$, $p < 0.001$); because this correlation describes agreement between two ways of measuring the same participants' trust, rather than the effect of explainability on trust, it is not included as an XAI-to-trust association above (see Methods).

Cross-study comparison emphasizes certain patterns that are highly notable in both synthesizing and contrasting information. First, there is a consistent and positive relationship between transparency level and trust across every study in the set (no source conclusively found that XAI harmed trust under standard conditions). Second, the failure-resilience finding from Hao et al. (2026), that XAI's greatest governance value may not be building initial trust but rather protecting existing trust when AI systems fail, is especially impactful for policy design: Hao et al. report that trust dropped significantly following a failed interaction in the non-explainable condition ($p < 0.001$), while trust in the explainable condition remained comparatively more stable. Third, Ahn et al.'s (2021) counterpoint emphasizes that interpretability alone does not reliably build behavioral trust and that outcome feedback is a more consistent method: across two large controlled experiments ($N = 800$ and $N = 711$), outcome feedback increased trust while global and local interpretability methods did not show a consistent or significant effect on their own. This suggests that XAI requirements work as a dual function that combines legal/ethical accountability and trust maintenance, rather than reliably driving trust gains on their own in every context. Fourth, the competence bias that Cheung and Ho (2025) identified, the public's tendency to trust AI developers based on technical ability rather than ethical conduct, shows how vulnerability in voluntary trust-building may depend on external oversight, since without it, the ethical dimensions of trust that voluntary frameworks are least likely to address on their own may go unchecked.

4. Discussion

This synthesis makes a direct and specific contribution to the growing literature on AI governance by combining and cross-comparing empirical findings from behavioral, organizational and survey-based research traditions into a single comparative framework. Previous reviews have treated transparency and governance

as mostly separate concerns that do not overlap in theory or practice. However, this study bridges the two foundational principles through quantitative and qualitative data suggesting that the mechanism linking XAI to public trust is not just technical but also institutional. The trust transfer finding from Li et al. (2021) evidently supports this concept, which is also further exemplified by the AI promoter dynamics from Li et al. (2021) and the governance framework of Li et al. (2024), suggesting that governance boards are not just requirements within procedural checklists but also trust assets whose credibility is transferable to the systems they monitor. These findings should be interpreted as associative rather than strictly causal, due to the mixed methodological nature of the included studies and because none of the six sources directly compares mandatory and voluntary governance or directly tests independent board authority. This framing has direct influence on how policymakers in the United States can craft feasible and realistic AI regulation that aims to directly address the trust deficiencies created by black-box AI systems.

The primary policy implication is that voluntary self-governance is simply not enough. Even the strongest transparency-based associations reported in **Table 1** and **Table 2** (Cheung and Ho's integrity pathway, $\beta = 0.28$; Li et al.'s AI promoter trust, $\beta = 0.532$) are considerably larger than the coercive/compliance-based association (directive leadership, $\beta = 0.129$), even though these values are reported descriptively here rather than pooled into group means. However, it is essential to consider context when evaluating self-governance, as it remains context-dependent and weak in the absence of enforcement. This evident weak point suggests that voluntary systems may improve short term trust signals, but mandatory systems are more likely to sustain institutional accountability over time. A government-centered, independently run governance board with mandatory XAI and outcome-reporting requirements could help close the accountability gap identified across the six sources through a structured approach. Specifically, the board's mandate should match the three-dimension framework proposed by Li et al. (2024), which addresses trustor-side deficits (in reference to AI literacy and public education), trustee-side deficits (fair auditing practices), and context-side deficits (legal accountability for algorithmic harm). The competence bias finding from Cheung and Ho further implies that audit criteria must go beyond technical performance metrics to include integrity and benevolence dimensions, which are both domains where XAI's associations with trust are largest and where voluntary frameworks typically have been the weakest. The failure resilience data from Hao et al. is directly relevant for AI developers and organizations. Taking advantage of the benefits of deploying XAI as a failure communication tool may be the most effective trust-preservation investment available. Even a qualitative reduction in trust damage following failure suggests consequences for user retention, reputational capital, and regulatory scrutiny worth taking seriously.

It is important to note that several limitations exist in regards to the generalizability of these findings. First, the six studies vary between each other on a large

scale when considering factors like design, population, and measurement instrument. The Singapore XAI Survey is geographically and culturally different from the U.S. context central to the policy hypothesis, and the China Manufacturing Study represents a single organization with a highly specific job demographic (88.3% male engineers). Pooling effect sizes across such differently designed studies would introduce methodological risk, which is why this synthesis reports effect sizes descriptively (Table 1 and Table 2) rather than as a single combined estimate, and does not compute a formal test of heterogeneity such as a Q or I^2 statistic; any apparent overall pattern here should be read as a description of the six sources rather than a statistically tested general claim. Second, because related outcomes drawn from the same sample (the three Singapore pathways in Table 2; the three China Manufacturing pathways in Table 1) are treated as within-study comparisons rather than independent data points, these specific comparisons ultimately rest on only two underlying samples, which limits how far the pattern can be generalized. Third, Ahn et al. was accessed primarily through its reported outcomes rather than full methodological documentation. This is concerning because there is uncertainty in how its results were coded and weighted compared to the other five sources. Fourth, this synthesis does not include any sources measuring long-term or longitudinal trust dynamics, which Li et al. (2024) identified as an inverted-U relationship (transparency can lower trust over time), suggesting that the largely cross-sectional sources reviewed here may overestimate the durability of XAI trust gains.

The most prevalent and immediate gap in the existing literature is the absence of U.S.-specific experimental or longitudinal data on XAI and governance board effects. Bullock et al. provides the best available snapshot of American public opinion, but it does not manipulate governance structure or XAI exposure as variables and only measures attitude toward them. A way to build off that and take a step forward would be to conduct a U.S.-based randomized experiment that presents participants with AI decision outputs under different levels of explainability and governance oversight type (voluntary vs. mandatory vs. board-certified), measuring both explicit trust and behavioral measures like willingness to act on AI recommendations. Second, the failure resilience finding suggests a replication study in a high-stakes public sector context like AI-assisted criminal sentencing or medical diagnosis. These types of environments have high consequences of trust miscalibration, and governance design choices have the most direct public impact. Third, as more literature accumulates on this topic, a future study with a larger and more homogeneous set of sources could attempt a formal statistical meta-analysis, including a proper heterogeneity test, building on the descriptive comparisons presented here. Finally, future research should try to operationalize the "board legitimacy as transferable trust asset" construct in an empirical manner, with a huge emphasis on whether citizens exposed to information about an independent AI governance board exhibit higher trust in AI systems certified by that board compared to identical systems without such certification.

5. Conclusion

The current issues in AI governance reflect the limitations of existing enforcement mechanisms. As AI works its way into healthcare, employment, education, and banking, the evidence reviewed in this paper suggests that voluntary frameworks alone do not offer sufficient enforcement to maintain a high standard of compliance. Ultimately, less compliance and more freedom for corporations to decide on AI deployment may contribute to a shift in priority towards profit-fulfillment and market competition over public responsibility. Concrete and centrally enforced ethical guidelines are associated, in the sources reviewed here, with not only the generation of accountability but the maintenance of it. A government-centered framework combined with independent corporate Algorithmic Review Boards offers a path toward consistent oversight and clear responsibility, though the six-source synthesis conducted in this paper is associative rather than a direct causal test of this claim, and does not directly compare mandatory government-centered governance against voluntary industry-led governance or directly test independent board authority. In order to construct a functioning domestic model, the United States can draw inspiration from international models that already outline how to use mechanisms such as risk-based regulation and mandatory audits. What the United States cannot afford to do is allow governance to remain rigid. Unless proper reforms are conducted to ensure long-term security, issues like bias, discrimination, and privacy violations will likely continue to persist. However, with sincere, dedicated effort to enforcement, the U.S. can not only mitigate public risk but also help set up future generations for responsible innovation.

Acknowledgements

I would like to express my deepest gratitude to Evan Selinger, Emily Hadley, and Janine Sharbaugh for their invaluable guidance and insightful feedback throughout the development of this work. Their expertise and thoughtful critiques were instrumental in shaping the final version of this manuscript. I am also grateful for their unwavering encouragement and support, which greatly enriched my research process.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Accountability Framework for Federal Agencies and Other Entities (2021). U.S. Government Accountability Office. https://www.gao.gov/assets/720/716110.pdf

Ahn, D., Almaatouq, A., Gulabani, M., & Hosanagar, K. (2021). Will We Trust What We Don't Understand? Impact of Model Interpretability and Outcome Feedback on Trust in Ai. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3964332

AI RMF Development (2021). National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development

Bietti, E. (2020). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 210-219. https://doi.org/10.1145/3351095.3372860

Bullock, J. B., Pauketat, J. V. T., Huang, H., Wang, Y., & Anthis, J. R. (2025). Public Opinion and the Rise of Digital Minds: Perceived Risk, Trust, and Regulation Support. *Public Performance & Management Review, 48,* 1357-1388. https://doi.org/10.1080/15309576.2025.2495094

Cheung, J. C., & Ho, S. S. (2025). The Effectiveness of Explainable AI on Human Factors in Trust Models. *Scientific Reports, 15,* Article No. 23337. https://doi.org/10.1038/s41598-025-04189-9

Choung, H., David, P., & Seberger, J. (2023). *A Multilevel Framework for AI Governance*. https://arxiv.org/pdf/2307.03198

Dastin, J. (2018). *Insight—Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women*. Reuters. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Davies, P. (2024). *Could the New EU AI Act Stifle genAI Innovation in Europe? A New Study Says It Could*. Euronews. https://www.euronews.com/next/2024/03/22/could-the-new-eu-ai-act-stifle-genai-innovation-in-europe-a-new-study-says-it-could

Edgerton, B., & Oktem, C. (2025). *Four Ways Boards Can Support the Effective Use of AI*. EY. https://www.ey.com/en_us/board-matters/four-ways-boards-can-support-the-effective-use-of-ai

Gambelin, O. (2024). *Responsible AI: Implement an Ethical Approach in Your Organization*. Kogan Page.

Hadley, E., Blatecky, A., & Comfort, M. (2024). Investigating Algorithm Review Boards for Organizational Responsible Artificial Intelligence Governance. *AI and Ethics, 5,* 2485-2495. https://doi.org/10.1007/s43681-024-00574-8

Hao, S., Teng, F., Hou, R., Zhang, L., Wu, H., & Qi, J. (2026). Explainable AI and Echo State Networks Calibrate Trust in Human Machine Interaction. *Scientific Reports, 16,* Article No. 1189. https://doi.org/10.1038/s41598-025-30899-1

ISACA (2025). *Press Release: AI Use Is Outpacing Policy and Governance, ISACA Finds*. https://www.isaca.org/about-us/newsroom/press-releases/2025/ai-use-is-outpacing-policy-and-governance-isaca-finds

Langenkamp, M., Costa, A., & Cheung, C. (2019). Hiring Fairly in the Age of Algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3723046

Li, J., Zhou, Y., Yao, J., & Liu, X. (2021). An Empirical Investigation of Trust in AI in a Chinese Petrochemical Enterprise Based on Institutional Theory. *Scientific Reports, 11,* Article No. 13564. https://doi.org/10.1038/s41598-021-92904-7

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing Trustworthy Artificial Intelligence: Insights from Research on Interpersonal, Human-Automation, and Human-AI Trust. *Frontiers in Psychology, 15,* Article ID: 1382693. https://doi.org/10.3389/fpsyg.2024.1382693

Loufek, B., Vidal, D., McClintock, D. S., Lifson, M., Williamson, E., Overgaard, S. et al. (2024). Embedding Internal Accountability into Health Care Institutions for Safe, Effective, and Ethical Implementation of Artificial Intelligence into Medical Practice: A Mayo Clinic Case Study. *Mayo Clinic Proceedings: Digital Health, 2,* 574-583.

https://doi.org/10.1016/j.mcpdig.2024.08.008

Madiega, T. (2019). *EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation*. European Parliamentary Research Service.

Mithal, M. (2022). *The Algorithmic Accountability Act*. Antitrust Magazine Online. https://www.americanbar.org/content/dam/aba/publications/antitrust/magazine/2022/august/algorithmic-accountability-act.pdf

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3784238

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B. et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44). ACM. https://doi.org/10.1145/3351095.3372873

Ruschemeier, H. (2023). AI as a Challenge for Legal Regulation—The Scope of Application of the Artificial Intelligence Act Proposal. *ERA Forum, 23,* 361-376. https://doi.org/10.1007/s12027-022-00725-6

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. https://ai.equineteurope.org/system/files/2022-02/ICA2014-Sandvig.pdf

Schuett, J., Reuel, A.-K., & Carlier, A. (2024). How to Design an AI Ethics Board. *AI and Ethics, 5,* 863-881. https://doi.org/10.1007/s43681-023-00409-y

Text-H.R.7532—118th Congress (2023-2024): Federal AI Governance and Transparency Act (2023). Congress.gov. https://www.congress.gov/bill/118th-congress/house-bill/7532/text

Treasury Board of Canada Secretariat (2021). *Algorithmic Impact Assessment Tool*. Government of Canada. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Ugwudike, P. (2021). AI Audits for Assessing Design Logics and Building Ethical Systems: The Case of Predictive Policing Algorithms. *AI and Ethics, 2,* 199-208. https://doi.org/10.1007/s43681-021-00117-5

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach. *Computer Law Review International, 22,* 97-112. https://doi.org/10.9785/cri-2021-220402

Wallen, E. (2025). *How Top Banks Build Credit Risk Models: An Expert Guide for Analysts*. C&R Software. https://blog.crsoftware.com/how-top-banks-build-credit-risk-models-an-expert-guide-for-analysts

Appendix A: Annotated Literature Matrix—AI Governance, Transparency, and Public Trust

The table below condenses the full coding matrix maintained during data extraction (see Methods) into its core columns: research design/sample, the enforcement level or oversight structure each source discusses, the key transparency-trust result extracted, and this paper's assessment of the source's relevance. All reported statistics match the Results section above.

Source	Design & Sample	Enforcement Level/ Oversight Discussed	Key Transparency-Trust Result	Assessment
Bullock et al. (2025)	Secondary analysis of the 2023 AIMS survey; N = 1099 nationally representative U.S. adults	Examines both voluntary ("soft") and mandatory ("strong") regulatory approaches; compares government trust vs. industry trust as predictors of support	Government trust positively predicts regulatory support (β = 0.12 - 0.23); industry trust negatively predicts it (β = −0.14 to −0.27); perceived risk is the strongest predictor (β = 0.49 - 0.59)	Empirical basis for the claim that trust in government, not industry, is the primary predictor of regulatory support; correlational survey, does not manipulate governance structure directly
Cheung & Ho (2025)	Cross-sectional survey, SEM (MPlus 8.3); N = 1002 stratified Singapore respondents	Examines trust in AI engineers (human developers) as an accountability node rather than a specific enforcement level	XAI → integrity β = 0.28 ($p < 0.001$); XAI → ability β = 0.15 ($p < 0.01$); XAI → benevolence β = 0.13 ($p < 0.001$); only ability significantly predicted downstream attitude (competence bias)	Basis for Table 2; three outcomes from the same sample are treated as a within-study comparison rather than three independent effect sizes
Li et al. (2021)—China Manufacturing	Cross-sectional field survey; N = 180 engineers (88.3% male), Chinese petrochemical enterprise	Compares an institutional/voluntary mechanism (management commitment), a coercive mechanism (directive leadership), and a human-intermediary mechanism (AI promoter trust)	AI promoter trust β = 0.532 ($p < 0.001$, dominant predictor); management commitment β = 0.192 ($p < 0.01$); directive leadership β = 0.129 ($p < 0.05$, weakest)	Basis for Table 1; strongest evidence in the six-source set for a trust-transfer mechanism; three outcomes from the same sample are treated as a within-study comparison
Ahn et al. (2021)	Two randomized, pre-registered web experiments; N = 800 and N = 711, general public	Not a governance/enforcement study; tests interpretability (global and local) vs. outcome feedback as trust-building mechanisms	Outcome feedback increased trust consistently across both experiments; global and local interpretability alone did not show a consistent or significant effect on trust	Key counterpoint to the assumption that interpretability alone drives behavioral trust; supports the dual-function XAI argument in the Discussion

Continued

Hao et al. (2026)	2 × 2 between-subjects lab experiment; N = 120 university participants plus simulated users; CIFAR-10/SQuAD datasets	Not a governance study; tests Grad-CAM/attention-based XAI and failure resilience	Trust higher with explanation (M = 5.92, SD = 0.43) than without (M = 5.12, SD = 0.51); explicit-implicit trust r = 0.78 (a measurement-agreement correlation, excluded from **Tables 1-2**); trust dropped after failure without explanation ($p < 0.001$)	Anchors the failure-resilience argument in the Discussion; the r = 0.78 correlation is correctly excluded from the XAI-to-trust synthesis (see Methods)
Li et al. (2024)	Systematic review/ theoretical synthesis (no primary data collection); cross-domain (healthcare, finance, civil services)	Proposes a three-dimension governance framework (trustor, trustee, context); the only one of the six sources to explicitly discuss the need for a third-party, expert-staffed oversight body	Identifies an inverted-U relationship between transparency and trust over time; accountability safeguards identified as the strongest institutional predictor of trust	Structures the proposed governance board's mandate in the Discussion; a review-level synthesis rather than an independent empirical data point