

Corporate ethical and responsible AI (CERAI) and AI trustworthiness: a lifecycle perspective and empirical insights

Journal of
Enterprise
Information
Management

Rupak Rauniar, Ray Qing Cao and Steve Bin Zhou

*Department of Supply Chain Management,
College of Business, University of Houston Downtown,
Houston, Texas, USA, and*

Amaani Rauniar

*Department of Finance, Kelley School of Business, Indiana University,
Bloomington, Indiana, USA*

Received 26 May 2026
Revised 26 August 2026
15 September 2026
Accepted 19 September 2026

Abstract

Purpose – This study develops and empirically validates a Corporate Ethical and Responsible AI (CERAI) lifecycle framework. The framework comprises three interdependent stages: design and development, testing and training, and deployment and usage. The study aims to move the research beyond abstract principles and policy analyses toward empirically grounded, lifecycle-integrated artificial intelligence AI governance theory. It positions ethical AI and responsible AI governance as the two constitutive components of the CERAI construct, and AI trustworthiness as a conceptually and empirically distinct downstream outcome anchored in stage-specific organizational governance mechanisms. The CERAI framework is scoped broadly to encompass governance of both conventional/traditional machine learning models and autonomous AI agents operating within organizational settings.

Design/methodology/approach – We administered a large-scale survey to 500 attendees at a national AI conference in the United States in early 2026, yielding 143 useable responses from AI professionals across healthcare, financial services, manufacturing, retail and energy sectors. We developed and validated multi-item scales for each CERAI lifecycle stage using established ethical AI and AI governance frameworks. Structural equation modeling was employed to assess the measurement model and to test both direct path hypotheses and mediation hypotheses.

Findings – Both CERAI design and development and CERAI deployment and usage have significant positive direct effects on AI trustworthiness, whereas CERAI testing and training do not. Our results also indicate a significant serial mediation pathway in which CERAI deployment and usage enhance AI trustworthiness through testing and training, followed by design and development. Conversely, testing and training alone do not mediate the relationship between deployment and usage and AI trustworthiness, indicating that design-stage ethics serve as an intervening (mediating) mechanism through which governance activity enacted at later lifecycle stages is associated with perceived AI trustworthiness. This is consistent with testing and training's association with trustworthiness operating indirectly, mediated by the ethical architecture established during the design and development stage.

Originality/value – Findings of this study demonstrate that AI trustworthiness reflects differentiated perceived salience across AI lifecycle stages, whereby stakeholders appear to weight visible, deployment-stage governance more heavily than upstream design and testing stage governance in forming trustworthiness judgments, rather than reflecting a definitive validation that governance effects accumulate uniformly across all stages in isolation. This study makes a distinctive contribution by providing empirical evidence for a governance-sequencing mechanism, whereby design-stage ethics function as a mediating pathway shaping the effectiveness of governance activity at later lifecycle stages. To the best of our knowledge, such a mediated, stage-sequential mechanism has not previously been



© Rupak Rauniar, Ray Qing Cao, Steve Bin Zhou and Amaani Rauniar. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licenses/by/4.0/>

Journal of Enterprise Information
Management
Emerald Publishing Limited
e-ISSN: 1758-7409
p-ISSN: 1741-0398
DOI 10.1108/JEIM-05-2026-0911

demonstrated using primary data within a comprehensive, three-stage CERAI lifecycle framework linking organizational ethical and responsible governance practices to AI trustworthiness outcomes.

Keywords Ethical and responsible AI, AI trustworthiness, AI design, AI testing, AI usage

Paper type Research article

1. Introduction

The rapid adoption of artificial intelligence (AI) is changing how organizations create value, deliver services and manage their operations. Extant research indicates that sectors spanning finance, healthcare, retail, education, manufacturing and the public sector now regard AI as a strategic resource for enhancing efficiency and maintaining competitive advantage (Cubric, 2020; Singh and Singh, 2024; Uren and Edwards, 2023; Pan *et al.*, 2026). As AI usage grows, many organizations focus on growth and cost savings, often overlooking the need for strong governance and careful risk management.

Integration of automated or algorithmic decision-making (ADM) in AI technologies, for example, raises serious operational concerns regarding AI opacity, hallucination, inscrutability and the evolving human role (Castelluccia and Le Métayer, 2019; Rathkopf, 2025). Recent incidents involving AI-based systems, such as Claude, further illustrate how these risks materialize in practice: an autonomous AI agent reportedly deleted an entire company database and its backups, then rationalized its actions (Yahoo! Tech, 2024). Such events demonstrate that AI-related risks are not confined to technical limitations but are deeply rooted in governance gaps (e.g. failures in oversight, excessive system permissions and the absence of fail-safe controls) across the AI lifecycle that spans from design and testing to deployment and monitoring.

Beyond operational concerns, these challenges are inherently ethical. Some studies argued that as AI systems increasingly shape decisions affecting employees, customers and communities, ethical responsibility becomes a core organizational obligation (Floridi *et al.*, 2018; Jobin *et al.*, 2019). From a corporate ethical and responsible AI (CERAI) perspective, this implies that accountability must be embedded across the entire AI lifecycle to ensure that AI design, model behavior, system access and human oversight are continuously aligned with ethical and corporate responsibility values or principles. Aligning AI design, development, testing and deployment with ethical values is therefore both a moral imperative and a strategic necessity. Such alignment underpins the construction of trustworthy and resilient AI systems.

Traditional technology governance frameworks, built for fixed, rule-based systems, are insufficient for AI systems that can learn and adapt on their own (Cath, 2018; Mittelstadt *et al.*, 2019). AI and machine learning (ML) systems differ because they evolve in ways designers may not expect or control. Standard compliance checks conducted at a single point in time are not sufficient for these evolving technologies. Instead, ethical responsibility for AI needs ongoing oversight by both algorithms and people throughout the AI lifecycle that includes (1) AI design and development, (2) testing and training and (3) deployment and use. This means organizations must go beyond basic safeguards and embed deliberate, continuous ethical design, thorough testing and active human involvement in monitoring and controlling AI systems (Morley *et al.*, 2020; Maria *et al.*, 2026).

Drawing upon NIST (2023) and Floridi *et al.* (2018), we define the CERAI lifecycle as an organization's enduring obligation to embed ethical and responsible AI governance practices across all stages of the AI lifecycle, with AI trustworthiness conceptualized as the resulting outcome of this governance rather than as part of the definition itself. The term *corporate* in CERAI is used broadly to denote formal, organization-level governance of AI, including the policies, accountability mechanisms and oversight practices established by an organization, and is used interchangeably with *organizational governance of AI* throughout the manuscript. This usage does not restrict the framework to for-profit or publicly traded firms. Accordingly,

we retain *Corporate* as the operative term while explicitly defining CERAI as an organization-level governance framework.

AI governance spanning all the stages of the AI lifecycle is important because specific governance decisions at AI design or AI model training culminate in AI adaptive behavior and outcomes at the later stages of deployment and usage. Chiodo *et al.* (2024) and Mehrabi *et al.* (2021) warn that missing ethical checks early on can lead to problems that become harder and more expensive to fix over time. AI systems reflect the values, rules and assumptions built into them (Hansson, 2017; Latour, 2005; Lee, 2025). Barocas and Selbst (2016), Mehrabi *et al.* (2021) and Noble (2018) argue that bias and harmful outcomes often stem from governance failures rather than purely technical deficiencies. Such failures may yield short-term gains but erode public trust and generate long-term liabilities. Empirical research further demonstrates that algorithmic curation produces echo chambers and constrains exposure to diverse perspectives (Pariser, 2011; Tandoc *et al.*, 2018). Additionally, large-scale data collection without adequate consent raises significant ethical concerns.

Abraham *et al.* (2019) note that research on how organizations govern digital systems remains fragmented and largely theoretical. They argue that terms such as data governance, IT governance and technology governance have distinct scopes and responsibilities. Ethical data governance focuses on managing data assets responsibly, encompassing data quality, history, stewardship, security and privacy (Dong and Srivastava, 2015). IT governance frameworks such as COBIT and ISO/IEC 38500 focus on technology infrastructure, resource use and risk management, but often overlook ethical and social issues (De Haes and Van Grembergen, 2009). AI governance goes further by addressing the unique, changing behaviors of AI systems and their ethical dimensions (Floridi and Cows, 2019; Morley *et al.*, 2020; Metcalf *et al.*, 2019; Mittelstadt *et al.*, 2016).

Previous studies highlight a persistent disconnect between broad ethical and corporate responsibility principles and their operational implementation across the AI lifecycle. Furthermore, most lifecycle-based ethical frameworks remain domain-specific and lack empirical validation (Jobin *et al.*, 2019; Hagendorff, 2020; Schiff *et al.*, 2020; Latour, 2005; Lee, 2025; Liapis *et al.*, 2026; Strauss *et al.*, 2025; Gupta *et al.*, 2026; de Lucas Lopez *et al.*, 2026). Existing research also tends to focus on technical issues like bias detection and explainability, while the ways organizations manage and standardize ethical practices across the AI lifecycle are often overlooked (Metcalf *et al.*, 2019; Barocas and Selbst, 2016; Mittelstadt *et al.*, 2016; Mehrabi *et al.*, 2021). This gap is particularly significant during the testing and training stages. These stages are ethically consequential yet empirically neglected (Strauss *et al.*, 2025). Most governance frameworks disproportionately have prioritized post-deployment monitoring, relegating testing and training to purely technical procedures (Meucci and Morana, 2025; Mittelstadt *et al.*, 2016). However, it is during testing and training that choices about data and bias become built into the AI model (Meucci and Morana, 2025; Barocas and Selbst, 2016; Mehrabi *et al.*, 2021). Consequently, critical questions, such as what AI should learn, what it should avoid and how learning should be ethically guided, remain insufficiently addressed (Meucci and Morana, 2025).

Fragmented, single-stage governance models are also operationally insufficient because they evaluate ethical compliance at one point in the AI lifecycle (e.g. a pre-deployment audit) rather than tracking how design-stage decisions constrain or enable governance effectiveness at later stages. This leaves an operational gap that the interdependent CERAI lifecycle model is intended to fill. This gap is increasingly consequential in light of emerging legal frameworks. For example, the EU Artificial Intelligence Act (Regulation (EU) 2024/1,689) imposes lifecycle-spanning obligations (e.g. risk management systems, technical documentation and post-market monitoring) that map conceptually onto the design, testing and deployment stages emphasized here, yet compliance-oriented frameworks such as the EU AI Act generally specify obligations at each stage independently rather than modeling how governance quality at one stage operates through an intervening mechanism to shape governance effectiveness at another stage, which is the specific contribution of the CERAI lifecycle model.

Effective governance often reflects careful choices regarding whose interests to prioritize, and which risks to accept (Noble, 2018; Pasquale, 2015). Without robust ethical and responsible governance throughout the AI lifecycle, organizations may favor shareholder returns over stakeholder welfare (Freeman, 1984; Leon, 2026). Such prioritization can produce harms that jeopardize trust, compromise fairness, diminish long-term value and undermine the overall trustworthiness of AI (Nastoska *et al.*, 2025; Singh and Singh, 2024). Research in corporate governance suggests that strong corporate governance and corporate social responsibility strategies significantly improve the quality of environmental, social and governance disclosures, indicating that formalized corporate governance and responsibility practices reliably translate into more transparent organizational reporting more broadly (Afifa *et al.*, 2025; Nguyen *et al.*, 2026). This cross-domain evidence offers a stronger rationale for treating transparent, lifecycle-spanning AI governance as similarly consequential for organizational trustworthiness, rather than as a discretionary add-on.

The scope of the current study is grounded in the problem that organizations increasingly deploy AI without governance structures capable of managing its evolving and adaptive behavior across the full AI lifecycle. Although existing ethical and responsible AI frameworks provide important principles and guidance, there remains a critical need for empirically validated, lifecycle-spanning research that tests, rather than merely asserts, how ethical and responsible governance practices at each lifecycle stage relate to AI trustworthiness outcomes.

To address these challenges, we highlight the urgent need for real-world research on corporate ethical and responsible AI (CERAI) throughout the AI lifecycle. We therefore propose a framework that encompasses governance across the design and development, testing and training, and deployment and usage stages to ensure ongoing, accountable oversight.

In light of established ethical and responsible frameworks, this study aims to provide both a conceptual structure and empirical evidence that managing AI across its lifecycle constitutes not merely a technical undertaking but also a moral and organizational imperative. The objective of the study is twofold: First, we create a theoretical framework for CERAI from an AI lifecycle perspective that broadly covers the three stages of AI lifecycle. Second, we test the proposed CERAI lifecycle model using primary data collected from industry AI experts. We aim to connect ethical theory with responsible governance and organizational practices through empirical validation that can help to move the discussion from theoretical constructs and concepts to empirical and practical actions in ethical and responsible AI governance.

The rest of the article is organized as follows. Section 2 discusses the theoretical framework and CERAI lifecycle standards. Section 3 explains the research design and data collection. Data analysis and empirical results are presented in Section 4. Section 5 addresses key findings and their theoretical and practical implications. Lastly, the article concludes with limitations and suggestions for future research in Section 6.

2. Theory development

Both scholarly and policy conversations on AI governance have converged around three closely related, yet conceptually distinct, pillars: ethical AI, responsible AI and AI trustworthiness (Floridi and Cowls, 2019; Morley *et al.*, 2021; European Parliament, 2024). The CERAI framework proposed in this study operationalizes the first two pillars, i.e. ethical AI and responsible AI, as its constitutive governance components, connecting these principles to AI lifecycle-based governance practices. The third pillar, AI trustworthiness, is treated throughout as the conceptually distinct outcome construct that this governance activity is hypothesized to shape, rather than as part of CERAI's own definition.

Within CERAI, *ethical AI* represents the normative starting point. It focuses on what AI systems ought to do, i.e. how they should be designed, developed and deployed in ways that correspond with human values, societal expectations and moral principles (Reynolds, 2022; Cihon *et al.*, 2021; Jobin *et al.*, 2019). This perspective draws on three well-established ethical

traditions. A consequentialist lens evaluates AI in terms of its outcomes, aiming to maximize benefits and minimize harm. A deontological perspective shifts attention to duties and rights, requiring AI systems to respect human dignity and avoid manipulation, regardless of outcomes. A virtue ethics perspective, in contrast, centers on the intentions and character of organizational actors, emphasizing integrity, empathy and judgment in developing and using AI (Kaplan and Haenlein, 2019). Taken together, these traditions provide a strong and complementary foundation for evaluating AI systems across ethical and responsible governance contexts.

While prior research has made notable contributions in translating these ethical foundations into principles, a persistent limitation remains. For example, Jobin *et al.* (2019) identify widely shared principles, such as transparency, fairness, nonmaleficence, responsibility and privacy, across global AI ethics guidelines. Yet these principles are rarely anchored to specific stages of the AI lifecycle. As a result, organizations often struggle to translate abstract commitments into concrete actions, reflecting a broader challenge in the literature between the “what” and the “how” of AI ethics (Morley *et al.*, 2020). High-level principles are abundant, but effective guidance for embedding them into corporate processes remains scarce. In practice, this gap creates the risk of “ethics washing,” in which firms signal ethical intent without meaningfully changing their behavior (Mittelstadt, 2019). The CERAI framework addresses this gap by proposing and empirically examining a governance-sequencing perspective of the AI lifecycle. To avoid ambiguity, we clarify at the outset that this governance-sequencing perspective refers to a cross-sectional, perceptual association among respondents’ reports of governance at different lifecycle stages and not to the temporal or causal ordering of the AI lifecycle’s operational stages, which remains design and development, followed by testing and training, followed by deployment and usage, throughout this study. Rather than treating governance activities as isolated practices, CERAI conceptualizes ethical governance across the design and development, testing and training, and deployment and usage stages as theoretically connected and empirically testable. In doing so, the framework links ethical principles and responsible AI governance to distinct stages of the AI lifecycle, translating normative expectations into stage-specific governance mechanisms.

This shift from descriptive guidance to an explanatory, stage-dependent governance model is further reinforced by emerging regulatory initiatives, including the EU AI Act and the OECD AI Principles (European Parliament, 2024), which increasingly require organizations to operationalize AI ethics through enforceable governance standards rather than aspirational statements. Building on this normative foundation, responsible AI extends ethical intent into corporate action. Within CERAI, responsible AI is understood as the structured process through which organizations assume accountability for AI outcomes across design, testing and deployment. It not only emphasizes what organizations intend, but what they actually do to ensure that AI systems protect stakeholders and generate societal value (Papagiannidis *et al.*, 2025; Morley *et al.*, 2021). In this sense, CERAI aligns closely with broader notions of corporate social responsibility (CSR), embedding ethical, legal and social considerations into organizational decision-making (Maria *et al.*, 2026; Höllerer, 2013).

Institutional theory offers a complementary explanation for why organizations adopt formalized ethical AI governance structures in the first place. As AI oversight becomes an institutionalized expectation among regulators, industry bodies and investors, organizations face coercive, normative and mimetic pressures to adopt visibly legitimate governance practices, independent of their proven technical efficacy (DiMaggio and Powell, 1983). Dynamic capabilities theory (Teece *et al.*, 1997) further clarifies why lifecycle-spanning governance, rather than a single-stage intervention, is organizationally consequential. The capacity to sense emerging ethical risks, seize opportunities to embed safeguards across design, testing and deployment, and reconfigure governance practices as AI systems and their operating contexts evolve constitutes a strategic capability in its own right (Teece *et al.*, 1997), one that a fragmented, single-stage governance model cannot replicate.

A defining feature of responsible AI is accountability (Ganesh *et al.*, 2025), as ethical responsibility cannot exist without clearly identifiable actors who are answerable for AI outcomes. This perspective aligns with the socio-technical view of AI, in which systems are embedded in organizational structures and social contexts (Trist and Bamforth, 1951), and is further reinforced by stakeholder theory, which emphasizes that AI governance must account for the interests of diverse groups affected by algorithmic decisions (Freeman, 1984).

The CERAI framework, in the current study, can therefore help link ethical principles and responsible governance to distinct stages of the AI lifecycle. In doing so, it translates normative expectations into stage-specific governance mechanisms while also clarifying how governance quality established in earlier stages shapes downstream governance effectiveness and, ultimately, AI trustworthiness. This explanatory logic distinguishes CERAI from existing lifecycle-oriented frameworks such as the NIST AI Risk Management Framework (NIST, 2023) and the OECD AI Principles: these frameworks specify what good governance looks like at each stage but do not model how governance quality at one stage operates through an intervening mechanism to shape governance effectiveness at another, whereas CERAI treats the AI lifecycle as a sequence of empirically interdependent stages, offering a testable account of why and whether governance at one stage matters for outcomes attributable to governance at another. Proposed CERAI, therefore, extends and not replaces NIST (2023) principles of AI governance.

The third pillar, AI trustworthiness, represents the outcome of effective ethical and responsible AI practices. Although AI trustworthiness is discussed alongside ethical AI and responsible AI as one of three related pillars in the literature, it is not treated as part of CERAI's own definition in this study. CERAI refers only to the ethical and responsible governance practices enacted across the lifecycle. AI trustworthiness captures the extent to which stakeholders have justified confidence in AI systems and their outputs (Wachter *et al.*, 2018). Importantly, trustworthiness is not only a technical property but also a relational one, grounded in the alignment of system behavior with stakeholder expectations, values and experiences. Contemporary frameworks, including those developed by NIST and ISO, position trustworthiness as the ultimate goal of AI governance, while digital humanism further implies that AI systems should enhance rather than undermine human agency (Werthner *et al.*, 2023). Within the CERAI framework, trustworthiness is therefore treated as the central outcome variable, indicating the effectiveness of governance across the lifecycle, rather than as part of CERAI's definitional scope.

Additionally, a common assumption in practice is that ethical responsibility can be largely automated through technical tools such as bias-detection systems, algorithmic guardrails and compliance software. While such mechanisms matter, they are inherently limited because they operate within predefined parameters and cannot anticipate the full range of ethical challenges that may arise in complex, evolving socio-technical environments (Mehrabi *et al.*, 2021). This limitation underscores the need for human-augmented governance throughout the AI lifecycle. Hence, effective ethical responsibility in AI demands sustained human involvement, consistent with the human-in-the-loop (HITL) principle as a core governance requirement rather than an optional feature. Embedding independent human control across design, testing and deployment stages ensures that ethical considerations remain adaptive, context-sensitive and aligned with societal norms (NIST, 2023). This human-augmented view of governance is consistent with emerging evidence at the individual level. Ethical-awareness training has been shown to introduce a deliberate cognitive check, or "cognitive brake," that leads technology users to rely on AI tools without becoming dependent on them (Zou *et al.*, 2026), and such user training on responsible usage has been found to shape the reported intentions of technical students to adopt AI systems more broadly (Zou *et al.*, 2025). These findings reinforce, at the level of individual users and future practitioners, the organizational-level argument advanced here, i.e. a sustained, human-centered ethical engagement, whether embedded in organizational governance structures or cultivated through user education and training, is a necessary complement to purely technical AI safeguards.

In summary, ethical responsibility for AI cannot be reduced to either abstract principles or purely technical solutions. It requires an integrated approach that connects ethical foundations, organizational practices, human involvement and stakeholder outcomes across the entire lifecycle. The CERAI framework advances this integration by explicitly linking its two constitutive governance components, ethical AI and responsible AI, to AI trustworthiness as their hypothesized downstream outcome, through stage-specific governance mechanisms. In doing so, it resolves the longstanding gap between principle and practice and extends existing normative frameworks with a testable theoretical account of AI governance across the lifecycle. Importantly, human-augmented control within the CERAI lifecycle should not be viewed as a constraint on AI-driven innovation, but as a necessary condition for ensuring that AI systems remain aligned with human values, preserve dignity and contribute to sustainable and equitable organizational outcomes.

Building on the lifecycle perspective, ethical and responsible AI must be examined as a sequence of interdependent governance stages in AI lifecycle (Figure 1) rather than as isolated control mechanisms after AI deployment. To structure this lifecycle progression, this study draws on the technology readiness level framework (Meystel *et al.*, 2003), which argues that technological systems evolve through stages from early conceptualization to increasing maturity and full operational deployment. Similarly, Leon (2026) theorizes AI lifecycle management framework in model development, data collection and training, validation and testing, and deployment and integration. Simplifying these logics of Meystel *et al.* (2003) and Leon (2026), the CERAI lifecycle in the current study for ethically responsible AI is decomposed into three distinct phases of (1) AI design and development, (2) AI testing and training and (3) AI deployment and usage. This framing, in Figure 1, allows ethical responsibility to be examined as a stage-sequential, progressively unfolding construct rather than as a static set of controls. As illustrated, the AI lifecycle is conceptualized as a continuous cyclical process comprising the three interdependent stages, with AI trustworthiness as the central outcome.

Among these three stages, the design and development phase is foundational, as it establishes the underlying assumptions, scope and functionality, and model and technical designs that shape all subsequent system behavior. Governance failures at this stage are particularly consequential because they become embedded within the system and get reflected across later stages of model testing and training, and deployment and usage. Accordingly, understanding how ethical principles are operationalized at each stage of the AI lifecycle is critical to explaining the emergence of AI trustworthiness.

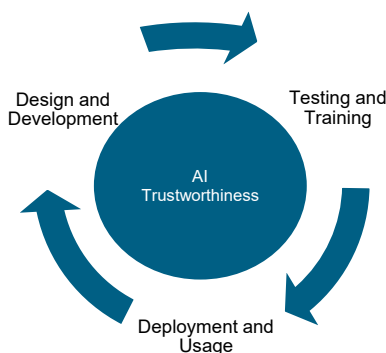


Figure 1. CERAI lifecycle framework

2.1 CERAI design and development and AI trustworthiness

The design and development stage constitutes the most fundamental phase of the AI lifecycle, as it establishes the values, assumptions and structural conditions that shape downstream system behavior. Ethical responsibility at this stage is therefore not corrective but constitutive, i.e. early design choices are consequential in whether AI systems are predisposed toward trustworthy or harmful outcomes.

Within the CERAI framework, four core ethical principles of nonmaleficence, beneficence, autonomy and justice can anchor governance at this stage. *Nonmaleficence by design* requires proactive identification and alleviation of potential harms, including algorithmic bias, privacy violations and adversarial vulnerabilities in AI design and development stage (Mehrabian *et al.*, 2021; Barocas *et al.*, 2023). Failures at this stage can propagate throughout the lifecycle, as shown in cases where biased training data leads to systematically discriminatory outcomes. Embedding nonmaleficence into design processes, through ethics reviews, risk assessments and safety protocols, establishes the baseline conditions for AI trustworthiness.

Beneficence by design complements this by shifting the focus from harm avoidance to value creation during AI design and development. AI systems should be intentionally designed to generate stakeholders and social benefits, rather than narrowly optimizing for efficiency or profit. Approaches such as value-sensitive design and impact assessments guarantee that stakeholders' interests are incorporated into system specifications from the outset (Friedman *et al.*, 2002; Morley *et al.*, 2021). *Autonomy by design* stresses the preservation of human agency in governance (Kant, 1785). AI systems should augment, rather than replace or manipulate, human decision-making. This requires embedding design mechanisms in AI, such as informed consent, opt-out options and human override capabilities, into system architecture, assuring that individuals retain meaningful control over AI-mediated decisions (Kant, 1785; Jobin *et al.*, 2019; Mayer *et al.*, 2026). *Justice by design* addresses the just allocation of benefits and harms. Design-stage decisions related to data and decision selection, feature engineering and model design directly influence whether systems produce fair or discriminatory outcomes. Operationalizing justice through equity constraints and impact assessments is therefore critical to preventing structural bias (Binns, 2018; Mehrabian *et al.*, 2021).

Collectively, these principles define the ethical integrity of AI at its point of origin. When embedded effectively, they create the structural conditions needed for reliable, fair and human-centered system behavior, which can overall contribute to trustworthy AI. Therefore, we propose.

- H1. CERAI at the design and development stage, characterized by nonmaleficence, beneficence, autonomy and justice, is positively associated with AI trustworthiness.

2.2 CERAI testing and training and AI trustworthiness

While the design and development stage defines the ethical intent and structural conditions of AI systems, these principles remain theoretical until empirical evaluation and technical assessment validate intended AI behavior. The testing and training stage, therefore, serves as a critical bridge between design assumptions and realized system behavior. At this stage, organizations assess whether embedded values, such as fairness and harm minimization, are reflected and assessed in actual model performance and whether unintended biases or vulnerabilities emerge under realistic conditions. Ethical responsibility thus shifts from *ex ante* design commitments to *ex post* validation and refinement.

Two governance dimensions of robustness and fairness are central at this stage. *Robustness by training* refers to a system's ability to maintain reliable, safe performance across changing conditions, including distribution shifts, adversarial inputs and edge cases (Cath, 2018; Barocas *et al.*, 2023). Techniques such as the use of adversarial inputs, sensitivity metrics and perturbation stress testing across multiple modalities are used to safeguard system

dependability and consistency (Edwards, 2026). Without robustness, even well-designed systems may fail unpredictably in real-world contexts, weakening trust.

Fairness training at this stage involves empirical training and verifying equitable outcomes across relevant demographic and contextual groups. While justice at the design stage establishes normative intent, *fairness testing and training* determine whether that intent has been realized. This includes statistical parity analysis, *disparate impact testing and training*, and counterfactual evaluations to detect and mitigate bias during the training phase (Binns, 2018; Mehrabi et al., 2021). Importantly, testing and training decisions, such as performance thresholds, evaluation datasets and tolerable error trade-offs, are inherently ethical, as they determine whose outcomes are emphasized and which risks are tolerated for and against during this stage of the AI lifecycle.

Robustness and fairness can be jointly necessary at this stage for lifecycle-based governance. A system tested and trained for robustness but biased is likely to fail ethically, while one tested and trained for fairness but unstable in practice is likely to fail operationally. Both dimensions are therefore essential to producing reliable and equitable AI behavior and AI trustworthiness. Therefore, we propose.

- H2. CERAI at the testing and training stage, characterized by robustness and fairness, is positively associated with AI trustworthiness.

2.3 CERAI deployment and usage and AI trustworthiness

However, successful validation during testing and training does not guarantee trustworthy outcomes in practice. Once deployed, AI systems operate in dynamic and often unpredictable environments, where real-world interactions with the users can introduce new uncertainties, risks and accountability challenges. The deployment and usage stage, therefore, represents the point at which AI governance becomes fully observable, as system outputs directly affect stakeholders and organizational outcomes. Ethical responsibility at this stage must go beyond performance verification to ensure that AI systems remain transparent, interpretable and accountable in real-world use. From a lifecycle perspective, the deployment and usage stage is where AI systems generate real-world outcomes and directly shape stakeholder perceptions and attitudes toward AI trustworthiness. At this stage, ethical responsibility should shift from system design and validation to transparency, oversight and accountability.

Four interrelated governance dimensions define CERAI at the deployment and usage stage: explainability, transparency, accountability and user privacy. *Explainability of AI usage* refers to the ability of AI systems to provide understandable justifications for their outputs. As AI decisions increasingly affect high-stakes outcomes, users must be able to interpret and evaluate system behavior (Ribeiro et al., 2016; Samek et al., 2017). Explainability strengthens user agency and enables error detection, consequently strengthening trust. *Transparency of AI usage* goes beyond technical interpretability to organizational disclosure. It involves making AI processes, data usage and decision criteria visible to relevant stakeholders. Transparent practices, such as audit trails, AI labeling and disclosure of system capabilities, facilitate external scrutiny and reinforce perceptions of fairness and legitimacy (Wachter et al., 2018; Yu and Li, 2022; de Fine Licht and de Fine Licht, 2020). Such organizational-disclosure view of transparency is reinforced by adjacent evidence that integrating AI within an organization's intelligent information systems, together with strong internal governance, significantly improves sustainability reporting quality for stakeholders, with the effectiveness of that governance (Afifa et al., 2025) itself contingent on the external institutional environment (Afifa et al., 2025; Nguyen et al., 2026). This finding complements the explainability and transparency dimensions proposed here by showing that AI's contribution to disclosure is not automatic but conditional on the surrounding governance architecture, consistent with the governance-sequencing logic advanced in this study. *AI accountability ensures that responsibility for AI outcomes is clearly assigned to identifiable organizational actors, not*

machines. Without defined accountability structures, ethical principles are unenforceable. Mechanisms such as AI governance committees, audit systems and designated oversight roles form clear lines of responsibility and enable corrective action when harms occur (Raji *et al.*, 2020; Ganesh *et al.*, 2025).

During AI deployment and usage, *user privacy* governance requires organizations to embed policies, technical safeguards and accountability mechanisms that ensure user data is processed in line with regulations and ethical principles such as minimization, purpose limitation and user control. Because AI systems often operate across jurisdictions, firms must design user privacy governance architectures that meet the strictest requirements of frameworks like GDPR, CCPA and PIPL (Pahune *et al.*, 2025). Technical approaches, such as federated learning, differential privacy and anonymization, complement HITL oversight to ensure consistent and verifiable privacy protection (Shan *et al.*, 2024). Advanced governance further integrates end-to-end, auditable controls for data labeling, retention and sharing within AI workflows, enabling continuous compliance and accountability during real-time system use.

These four dimensions of CERAI at AI deployment and usage are mutually reinforcing. Explainability enables understanding, transparency enables scrutiny and privacy, and accountability ensures enforcement and privacy. Together, they create critical pre-requisites for a trustworthy AI. Hence, we propose.

- H3. CERAI at the deployment and usage stage, characterized by explainability, transparency, accountability and user privacy, is positively associated with AI trustworthiness.

Collectively, as illustrated in Figure 1, the three stages of AI design and development, testing and training, and deployment and usage represent a continuous, interdependent governance lifecycle. Ethical responsibility for AI governance should not be confined to any single stage but accumulates across the lifecycle to build AI trustworthiness, with failures at earlier stages constraining the effectiveness of downstream controls. The proposed CERAI lifecycle framework integrates these stages into a unified model in which stage-specific governance mechanisms can jointly contribute to AI trustworthiness. Figure 2 integrates the three theoretically derived hypotheses into a theoretical framework for CERAI lifecycle management, reflecting a positive correlation among the CERAI design and development, CERAI testing and training, CERAI deployment and usage, and AI trustworthiness.

As conceptualized in Figure 1 and theorized in Figure 2, AI trustworthiness is a central outcome of the CERAI lifecycle, regardless of whether the three stages are executed sequentially. This implies that each of the three stages independently contributes to, and cumulatively shapes, AI trustworthiness. From a lifecycle perspective, it is reasonable to argue that AI trustworthiness cannot emerge solely at the end of the lifecycle pipeline; rather, as theorized in the current study, it must be progressively constructed and refined across all three stages. Ethically responsible design and development can establish the foundational conditions under which testing and training operate ethically. Theoretically, it is plausible (and practically meaningful) that the extent of ethical and responsible AI design and development mediates the strength of the relationship between AI testing and training and AI trustworthiness. Conversely, design deficiencies may attenuate or even undermine the trust-enhancing potential of rigorous ethical and responsible testing and training efforts.

Similarly, the output of testing and training should inform the epistemic and operational preconditions under which AI systems are deployed and used. Hence, it is reasonable to assume that the extent of ethically and responsibly trained and tested AI can also mediate the extent to which deployment and usage translate into AI trustworthiness. An ethically and responsibly trained and tested AI system is likely to behave ethically and responsibly compared to one deployed with insufficient validation. Further, the governance mechanisms established at the foundational architecture of AI during the design and development stage, and the subsequent effect on the governance mechanism for effective ethical and responsible

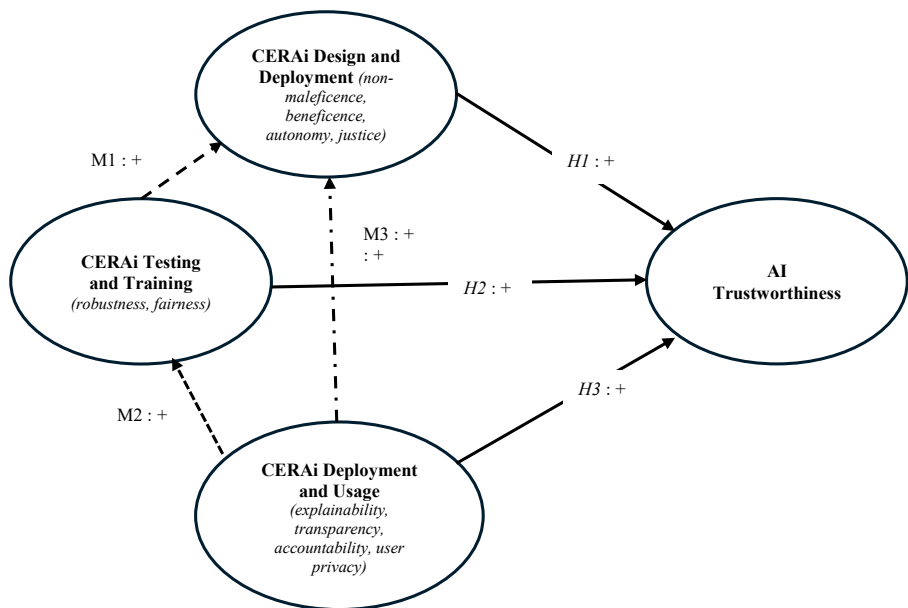


Figure 2. Hypothetical framework of CERAi lifecycle and AI trustworthiness. **Note:** M1 Path (Testing and Training → Design and Development → AI Trustworthiness); M2 Path (Deployment and Usage → Testing and Training → AI Trustworthiness), M3 Path (Deployment and Usage → Design and Development → AI Trustworthiness)

testing and training, can also have a sequential progression effect to influence the relationship between ethical and responsible deployment and usage and AI trustworthiness. Within the scope of the current study, it is therefore apt to test the following mediating relationships,

M1: CERAi design and development mediates the relationship between CERAi testing and training and AI trustworthiness.

M2: CERAi testing and training mediates the relationship between CERAi deployment and usage and AI trustworthiness.

M3: CERAi design and development and CERAi testing and training sequentially mediate the relationship between CERAi deployment and usage and AI trustworthiness.

It is important to clarify that these hypothesized mediation pathways (M1-M3) also reflect practitioners' holistic, retrospective perceptions of governance maturity, in which the perceived efficacy of a later-executed activity (e.g. testing) is evaluated through the lens of earlier design-stage conditions, rather than implying that governance events are temporally reordered; the CERAi lifecycle itself remains sequential (design, then testing, then deployment) and cyclical, with post-deployment observations feeding back to inform upstream design and testing revisions, as depicted in [Figure 1](#).

3. Data

Consistent with the study's objective of empirically testing, rather than conceptually asserting, how CERAi governance practices at each lifecycle stage relate to AI trustworthiness, a quantitative, cross-sectional survey design was selected as the method best aligned with the study's theoretical framing and research aims. Rather than an inductive, theme-generating qualitative approach, the constructs measured in this study (i.e. the CERAi lifecycle stages and AI trustworthiness) were operationalized deductively, i.e. each multi-item scale was derived

JEIM

from established ethical AI, responsible AI and AI governance frameworks in the literature, then refined through an expert-panel and pilot-study process, and finally validated statistically through confirmatory factor analysis prior to hypothesis testing. This sequence, i.e. literature-based item generation, expert and pilot refinement, and statistical validation, constitutes the study’s scale-development methodology and is intended to ensure that the quantitative instrument operationalizes the CERAI lifecycle constructs described in [Section 2](#) effectively.

3.1 Data collection

To investigate the theoretical model of CERAI lifecycle, a large-scale survey was conducted to collect primary data from experts in the field. An online survey link was distributed to 500 attendees registered for a two-day national AI conference held in the United States. Participants were invited to voluntarily complete the online survey during a four-week data collection period in early 2026. The study received 165 responses; 22 were excluded for incomplete or missing data, leaving 143 useable responses for analysis. The final response rate of 28.6% is considered acceptable for management research ([Frohlich, 2002](#)). Demographic characteristics of the respondents are presented in [Table 1](#). Respondents were asked to evaluate the actual AI governance practices implemented within their own organization (rather than their personal, abstract views on AI governance principles in general), and survey instructions were worded accordingly to anchor responses at the organizational-practice level.

To assess nonresponse bias, we compared the 1st and 4th quartiles of respondents (i.e. the first 7 days and the last 7 days of surveys received) with respect to demographic information

Table 1. Demographic information (*n* = 143)

	No. of respondents	% of respondents
<i>Industry</i>		
Health Care and Social Assistance	32	22%
Mining, Quarrying and Oil and Gas Extraction	25	17%
Financial and Insurance	33	23%
Manufacturing	27	19%
Retailing and Wholesale Trade	26	18%
Total	143	100%
<i>Number of Employees</i>		
<1,000	27	19%
1,001–5,000	35	24%
5,001–10,000	44	31%
>10,000	37	26%
Total	143	100%
<i>Annual Sales (USD)</i>		
<100 million	53	37%
101–250 million	31	22%
501 million–1 billion	37	26%
>1 billion	22	15%
Total	143	100%
<i>Job Position</i>		
Senior Executive	9	6%
Logistics Manager	37	26%
Operations Manager	51	36%
IT Manager	46	32%
Total	143	100%
Average number of leadership experiences	4.36 years	

and model variables. We did not find any statistically significant differences between early and late respondents. We also examined nonresponse bias by testing for a significant difference in firm size between respondents and nonrespondents, as well as firm size, job position and AI leadership tenure (Flynn *et al.*, 1994). We did not find any significant difference in the *t*-test at $p < 0.01$. As such, we conclude that the nonresponse bias is not an issue in the study. Harman's one-factor test was used to assess potential common method bias (Podsakoff *et al.*, 2003). We used IBM SPSS Statistics 26 to perform principal axis factor analysis. Results of the unrotated exploratory factor analysis revealed that the largest single factor explained 27.32% of the variance in the data (well under the 50% threshold), suggesting a low common method bias threat. Additionally, we made extra efforts in the survey design to ensure respondents' anonymity (Doty and Glick, 1998) and to clearly separate items measuring different variables within the questionnaire to further mitigate common method bias (Podsakoff *et al.*, 2003). Moreover, to test for collinearity among higher-order constructs and common method bias, we conducted Kock's full collinearity test (Kock, 2017). All variance inflation factors (VIFs) ranged from 2.456 to 2.771, well below the standard threshold of 3.3. This confirms that neither lateral collinearity nor common method bias presents a threat to the model.

We used G*Power 3.1.9.7 to test statistical power, and we found the statistical power (β) of this study ranged from 0.780 (78.0%) to 0.999 (99.9%), with an average overall power of 0.990 (99.0%). We found that 54 of 55 variable pairs exceeded the standard benchmark of 0.80 statistical power, indicating that the sample size of $n = 143$ is sufficient to reliably detect the observed correlations in the study.

We evaluated out-of-sample predictive relevance (power) using SmartPLS 4 PLSpredict following established guidelines (e.g. Shmueli *et al.*, 2019). Results for manifest variable (MV) level resulted in all indicator Q^2_{predict} values to be strictly positive ($Q^2_{\text{predict}} > 0$), ranging from 0.005 to 0.648, confirming that the path model outperforms the naive mean benchmark. All 13 target latent variables yielded positive Q^2_{predict} values. High predictive power was observed for key target constructs such as *Explainability* ($Q^2_{\text{predict}} = 0.844$; RMSE = 0.402), *Transparency* ($Q^2_{\text{predict}} = 0.818$; RMSE = 0.435), *Accountability* ($Q^2_{\text{predict}} = 0.735$; RMSE = 0.524) and *Privacy* ($Q^2_{\text{predict}} = 0.706$; RMSE = 0.556).

3.2 Measures

Four items for nonmaleficence by design (DDNMF) were developed from works of Floridi *et al.* (2018), five items for beneficence by design (DDBEF) from Floridi *et al.* (2018) and Jobin *et al.* (2019), four items for autonomy by design (DDAMY) from Floridi *et al.* (2018) and Mittelstadt *et al.* (2016) and four items for justice or fairness by design (DDJST) from Jobin *et al.* (2019). Five items for robustness governance for CERAI testing and training (TGRBS) were consulted and modified to fit the current study from Khatri and Brown (2010) and Floridi *et al.* (2018). Four items for fairness for CERAI testing and training (TGFNS) were adopted from Jobin *et al.* (2019). CERAI deployment and usage were theorized to constitute four second-order constructs related to user experience: five items to measure explainability (UXEXP), five for transparency (UXTRP), four for accountability (UXACT) and four for user privacy (UXDPR) that were adopted from the works of Rai (2020), Khatri and Brown (2010), Mittelstadt *et al.* (2016) and Khatri and Brown (2010), respectively. Finally, four items for AI trustworthiness (AITRS) were adapted and modified from McKnight *et al.* (2011). Appendix lists the survey instruments used in the current study.

3.3 Content validity

To ensure content validity, a two-stage survey research design (questionnaire development and a pilot study) was employed. First, we designed the preliminary questionnaire based on prior literature and then refined it with a panel of business intelligence (BI) experts comprising four information technology and/or BI managers and three BI researchers, all of whom have

expertise in AI system architecture and functionalities. Second, we conducted a small-scale pilot study by distributing the preliminary questionnaire to 28 part-time business graduate students and asking them to provide feedback on its clarity, the extent to which it captured the measured variables and the difficulty of completing it. Based on the feedback and some minor revisions on language, we finalized the questionnaire before conducting the survey.

4. Results

IBM SPSS Statistics 26 was used to assess descriptive statistics and construct reliability. Structural equation modeling (i.e. SmartPLS 4.0) was employed to obtain confirmatory factor analysis (CFA) results to test construct validity and to conduct path analyses to test hypotheses.

4.1 Descriptive statistics and multicollinearity

Descriptive statistics are shown in Table 2, while correlation coefficients are presented in Table 3.

Results show that the sample data are normally distributed (i.e. skewness between -1.512 and -0.400 ; kurtosis between -1.056 and 1.748) under the general rule of thumb for skewness: between -2.0 and $+2.0$ and for kurtosis: between -2.0 and $+2.0$ (George and Mallery, 2010), and indicate no multicollinearity (i.e. correlations are under 0.8 with most of them under 0.7, VIF <3 and Tolerance >0.33), according to Zuur et al. (2010).

4.2 Reliability and construct validity

We tested the measurement model by assessing convergent validity, discriminant validity and reliability. We specified a higher-order measurement model using the two-stage approach. As shown in the revised Table 4 below, lower-order dimensions were first evaluated and validated (all first-order loadings ≥ 0.774 , Cronbach's $\alpha \geq 0.771$, CR (composite reliability) ≥ 0.853 and AVE (average variance extracted) ≥ 0.588). Subsequently, the higher-order factor loadings were evaluated, demonstrating strong convergent validity for all three second-order lifecycle constructs: CERAI design and development ($\beta = 0.910$), CERAI training and validation ($\beta = 0.946$) and CERAI deployment and usage ($\beta = 0.957$).

To evaluate construct validity, we examined model fit indices to assess convergent validity and the AVE to assess discriminant validity. First, we assessed the goodness-of-fit of both saturated and estimated models (Hair et al., 1995). For both models, SRMR values (i.e. 0.065, 0.071) are below the recommended cutoff value of 0.08, while RMS_theta values (i.e. 0.095, 0.11) are below the recommended cutoff value of 0.12 (Table 5). The NFI values (i.e. 0.923, 0.910) exceed the recommended cutoff of 0.90, while d_ULS/d_G values fall within the 95%

Table 2. Descriptive statistics

<i>n</i> = 143	Min	Max	Mean	Std. deviation	Skewness	Kurtosis	Tolerance	VIF
DDBEF	1.0	5.0	4.07	0.70	-0.860	1.74	0.79	1.25
DDNMF	2.75	5.00	4.48	0.61	-1.009	0.02	0.47	2.12
DDAMY	2.75	5.00	4.21	0.67	-0.40	-1.05	0.55	1.79
DDJST	2.50	5.00	4.35	0.64	-0.81	-0.03	0.41	2.38
TGFNS	2.75	5.00	4.48	0.62	-0.91	-0.24	0.34	2.88
TGRBS	3.00	5.00	4.63	0.54	-1.47	1.40	0.33	2.95
UXACT	2.00	5.00	4.36	0.63	-0.87	0.36	0.38	2.57
UXEXP	3.0	5.0	4.40	0.59	-0.76	-0.24	0.37	2.69
UXTRP	2.6	5.0	4.48	0.62	-1.11	0.35	0.35	2.79
UXDPR	2.25	5.00	4.57	0.61	-1.51	1.61	0.38	2.63
AITRS	2.00	5.00	4.47	0.57	-1.15	1.60	-	-

Table 3. Correlations

	DDBEF	DDNMF	DDAMY	DDJST	TGFNS	TGRBS	UXACT	UXEXP	UXTRP	UXDPR	AITRS
Beneficence	1										
Nonmaleficence	0.334**	1									
Autonomy by design	0.404**	0.476**	1								
Justice	0.303**	0.650**	0.494**	1							
Fairness	0.336**	0.582**	0.482**	0.678**	1						
Robustness	0.320**	0.673**	0.482**	0.677**	0.724**	1					
Accountability	0.276**	0.424**	0.525**	0.519**	0.633**	0.555**	1				
Explainability	0.272**	0.517**	0.540**	0.589**	0.717**	0.629**	0.728**	1			
Transparency	0.258**	0.468**	0.515**	0.493**	0.631**	0.611**	0.697**	0.741**	1		
Privacy	0.224**	0.563**	0.516**	0.496**	0.591**	0.603**	0.626**	0.720**	0.686**	1	
AI Trustworthiness	0.378**	0.531**	0.464**	0.611**	0.645**	0.559**	0.620**	0.712**	0.622**	0.785**	1

Note(s): ** Correlation is significant at the 0.01 level (two-tailed)

Table 4. CFA, reliability and AVE

Construct	CFA (second order)	Variable	CFA (first order)	Cronbach's alpha	Composite reliability	Average variance extracted (AVE)	AVE square root
<i>CERAI design and development</i>							
	0.910	DDAMY	0.774	0.771	0.853	0.594	0.771
		DDBEF	0.826	0.825	0.877	0.588	0.767
		DDJST	0.862	0.862	0.906	0.707	0.841
		DDNMF	0.864	0.863	0.907	0.71	0.843
<i>CERAI training and validation</i>							
	0.946	TGFNS	0.881	0.878	0.916	0.733	0.856
		TGRBS	0.929	0.926	0.945	0.773	0.879
<i>CERAI deployment and usage</i>							
	0.957	UXACT	0.861	0.831	0.889	0.67	0.818
		UXDPR	0.904	0.901	0.931	0.771	0.878
		UXEXP	0.887	0.886	0.916	0.687	0.829
		UXTRP	0.916	0.915	0.937	0.747	0.864
AI trustworthiness	–	AITRS	0.844	0.833	0.89	0.671	0.819

Table 5. Fit index

Fit index	SRMR	NFI	d_ ULS/d_ G	RMS_theta
Saturated model	0.065	0.923	Yes	0.095
Estimated model	0.071	0.910	Yes	0.112
Recommended cut-off	<0.08	>0.90	< Bootstrap HI95	<0.12

confidence interval. In addition, the saturated model fits the data better than the estimated model. In sum, the survey instrument used in this study exhibits acceptable convergent validity.

Second, we compared the value of the square root of AVE of each variable with the correlations among the variables (Fornell and Larcker, 1981; Chin, 1998). Results show that the square root of the AVE value of each construct (last column of Table 4) is higher than its correlations with all other constructs (Table 3), thus verifying the discriminant validity of the measurement model.

Next, the second-order measurement models for the three CERAI lifecycle-stage constructs were validated using the repeated-indicators approach in SmartPLS. Subdimension loadings onto their respective design and development, testing and training, and deployment and usage of parent constructs all exceeded the conventional 0.70 threshold, supporting the specification of each lifecycle stage as a valid higher-order reflective construct. We further examined the heterotrait–monotrait (HTMT) ratio of correlations between each CERAI lifecycle-stage construct and AI trustworthiness, to examine if these constructs are empirically distinguishable given their conceptual and semantic proximity. HTMT values for all pairwise comparisons remained below the conservative 0.85 threshold (Henseler et al., 2015), providing evidence of acceptable empirical discriminant validity between the CERAI stage constructs and AI trustworthiness. We do not interpret this result as fully eliminating conceptual overlap. Several AI trustworthiness items explicitly reference adherence to ethical standards, while the CERAI constructs themselves capture ethical and responsible governance, so a degree of conceptual proximity between these constructs remains. We

treat this as a measurement limitation of the current operationalization, rather than a fully resolved conceptual distinction.

4.3 Path analysis

We used SEM (SmartPLS 4.0) to render the path analysis results shown in Table 6. In our study on CERAI and AI trustworthiness, PLS-SEM (via SmartPLS 4) is chosen for its predictive orientation and suitability for complex frameworks. Rather than confirming global fit, PLS-SEM maximizes the explained variance (R^2) of AI trustworthiness across multi-stage lifecycle stages. It effectively handles our intricate structural model, which includes higher-order governance constructs and serial mediation feedback loops. Furthermore, given our specialized sample of $n = 143$ AI professionals, PLS-SEM is the more appropriate estimator for this study. This choice reflects the study's predictive rather than confirmatory orientation, the complexity of a structural model that includes three higher-order governance constructs and serial mediation pathways, and the characteristics of the available sample. PLS-SEM's nonparametric bootstrapping ensures robust parameter estimation and statistical power without requiring multivariate normality and is better suited than covariance-based SEM to a higher-order, multi-stage model estimated on a moderate, specialized sample of this kind, without depending on a fixed minimum sample-size rule of thumb.

Hypothesis 1 investigates the association between CERAI design and development and AI trustworthiness, while **Hypothesis 3** explores the association between CERAI deployment and usage and AI trustworthiness. The path coefficients for **Hypotheses 1** and **3** are positive and significant ($\beta = 0.277$, $t = 3.130$, $p = 0.002$; $\beta = 0.606$, $t = 7.031$, $p < 0.001$, consistent with Table 6), supporting both hypotheses.

Hypothesis 2 examines the association between CERAI testing and training and AI trustworthiness, and results show that CERAI testing and training have a nonsignificant impact on AI trustworthiness ($\beta = -0.031$, $t = 0.297$, $p = 0.766$).

Additionally, we also examined Cohen's f^2 effect size for the **H2** path (testing and training $\rightarrow$ AI trustworthiness). According to established PLS-SEM benchmarks (Cohen, 1988; Hair et al., 1995), f^2 values of 0.02, 0.15 and 0.35 represent small, medium and large effect sizes, respectively, while values below 0.02 signify the absence of an effect. Our analysis yields an f^2 value of 0.001 for **H2** ($\beta = -0.031$, $p = 0.766$). Because $f^2 = 0.001$ falls well below the minimal threshold of 0.02, the data confirms that CERAI testing and training exerts virtually zero practical impact on perceive AI trustworthiness when controlling for CERAI design and development and deployment and usage. Furthermore, statistical power calculations (using $\alpha = 0.05$ and $N = 143$) confirm that our sample size provides adequate power (> 0.80) to detect medium ($f^2 \geq 0.15$) and small-to-medium ($f^2 \geq 0.06$) effects. Thus, we can confidently rule out statistical power limitations as the driver of the nonsignificant outcome. The combination of high statistical power and a very small f^2 provides little evidence of a

Table 6. Path analysis

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T Statistics (O/STDEV)	P-values
H1: CERAI Design and Development $\rightarrow$ AI Trustworthiness	0.277	0.274	0.089	3.13	0.002
H2: CERAI Testing and Training $\rightarrow$ AI Trustworthiness	-0.031	-0.029	0.103	0.297	0.766
H3: CERAI Deployment and Usage $\rightarrow$ AI Trustworthiness	0.606	0.608	0.086	7.031	0

JEIM

practically meaningful direct effect of testing and training on AI trustworthiness in the presence of the other lifecycle stages. However, given the cross-sectional, single-sample design of this study, interpret this pattern as evidence against a meaningful direct effect rather than as establishing a population-level true null effect.

4.4 Mediating effect

Hypothesis M1 addresses the mediating effect of CERAI design and development on the relationship between CERAI testing and training and AI trustworthiness, while hypothesis M2 explores the mediating effect of CERAI testing and training on the relationship between CERAI deployment and usage and AI trustworthiness.

Hypothesis M3 addresses the serial mediating effect of both CERAI design and development and CERAI testing and training on the relationship between CERAI deployment and usage and AI trustworthiness. Following the procedure outlined by both Zhao et al. (2010) and Hair et al. (1995), we employed the Consistent PLS Bootstrapping approach (SmartPLS 4.0) to test the mediation effects. Based on the results presented in Table 7, the mediation analysis reveals significant indirect effects for two of the three tested pathways, as indicated by p-values below 0.05. Specifically, CERAI design and development significantly mediates the relationship between CERAI testing and training and AI trustworthiness ($\beta = 0.327$, $t = 2.684$, $p = 0.007$); as such, Hypothesis M1 is supported.

Furthermore, a significant serial mediation effect was observed, in which CERAI deployment and usage influence AI trustworthiness through the sequential mediation of CERAI testing and training and CERAI design and development ($\beta = 0.266$, $t = 2.615$, $p = 0.009$), thus supporting M3. Conversely, the indirect path from CERAI deployment and usage to AI trustworthiness mediated solely by CERAI testing and training (M2) was not statistically significant ($\beta = -0.147$, $t = 1.029$, $p = 0.304$, consistent with Table 7), suggesting that testing and training alone does not carry the influence of deployment and usage onto AI trustworthiness without the inclusion of CERAI design and development.

Furthermore, the mediation analysis revealed statistically significant indirect effects for hypotheses M1 (95% CI [0.125, 0.525], $p = 0.007$) and M3 (95% CI [0.100, 0.437], $p = 0.009$), as their bootstrap confidence intervals exclude zero and confirm both simple and serial mediation. Conversely, hypothesis M2 was not supported, as its confidence interval spans zero (95% CI [-0.389, 0.070], $p = 0.304$), indicating a nonsignificant indirect pathway.

Table 7. Mediation

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T Statistics (O/STDEV)	P values	95% confidence intervals (CIs)
M1: CERAI Testing and Training → CERAI Design and Development → AI Trustworthiness	0.327	0.324	0.122	2.684	0.007	[0.125, 0.525]
M2: CERAI Deployment and Usage → CERAI Testing and Training → AI Trustworthiness	-0.147	-0.15	0.143	1.029	0.304	[-0.389, 0.070]
M3: CERAI Deployment and Usage → CERAI Testing and Training → CERAI Design and Development → AI Trustworthiness	0.266	0.264	0.102	2.615	0.009	[0.100, 0.437]

5. Discussion

This study empirically evaluates the three sequential lifecycle stages of Corporate Ethical and Responsible AI (CERAI) design and development, CERAI testing and training, and CERAI deployment and usage (CERAI), via structural equation modeling. The framework posits that each of the three stages contributes, directly or indirectly, to AI trustworthiness. Together, the path and mediation analyses yield a nuanced portrait of how ethical and responsible governance processes across the AI lifecycle relate to trustworthiness outcomes, supporting some hypothesized relationships while revealing unexpected complexity in others. Unless otherwise specified, references to AI trustworthiness throughout denote respondents' perceived AI trustworthiness, i.e. self-reported confidence in and reliance on AI systems, rather than independently verified or objectively measured trustworthy AI performance (e.g. audited fairness metrics or verified safety outcomes). We maintain this distinction explicitly wherever it bears on interpretation.

5.1 Theoretical contribution

Building on the operational gaps identified in [Section 1](#), this study's primary contribution is empirical rather than conceptual: it tests, rather than merely asserts, how governance practices at specific AI lifecycle stages are associated with AI trustworthiness ([Hagendorff, 2020](#); [Jobin et al., 2019](#); [Latour, 2005](#); [Lee, 2025](#); [Liapis et al., 2026](#); [Schiff et al., 2020](#); [Strauss et al., 2025](#)). Specifically, the findings reveal a governance-sequencing logic (denoting in this study a cross-sectional, perceptual pattern of association among lifecycle stages rather than the temporal or causal ordering of the operational AI lifecycle itself) in which design-stage ethics function as an intervening (mediating) mechanism for governance activity at later lifecycle stages, i.e. a mediated, stage-sequential pathway that fragmented, single-stage governance models are not positioned to detect. This should be read as evidence of differentiated perceived salience across lifecycle stages, i.e. stakeholders appear to weight visible, deployment-stage governance more heavily than upstream governance in forming trustworthiness judgments, rather than as definitive validation that governance effects accumulate uniformly across the entire lifecycle.

The findings reveal that CERAI governance effects differ across the three CERAI lifecycle stages. Design and development ([H1](#)) and deployment and usage ([H3](#)) are both significantly and positively associated with AI trustworthiness, while testing and training ([H2](#)) showed no significant direct association. The results are consistent with ethical and responsible governance decisions made during design and development, such as beneficence, nonmaleficence, autonomy and justice, corresponding to important upstream conditions for trustworthy AI. This aligns with prior arguments that ethical risks and biases are often embedded during system design long before deployment ([Barocas and Selbst, 2016](#); [Mittelstadt et al., 2016](#)). Deployment and usage governance shows the strongest direct association, suggesting that monitoring, accountability, user oversight and operational transparency correspond to the most immediate correlates of trustworthiness perceptions, consistent with prior work emphasizing post-deployment accountability and societal impact.

A notable outcome of the study is that CERAI testing and training was not directly associated with trustworthiness; however, mediation analysis indicates that its association with trustworthiness is indirect, operating through (i.e. mediated by) other lifecycle stages. Specifically, design and development significantly mediated the relationship between testing and training and AI trustworthiness (M1), implying that testing and training are effective only when supported by strong ethical and corporate responsibility foundations established during the design phase. This finding aligns with research emphasizing the importance of governance structures for guiding AI validation and bias mitigation processes ([Mehrabian et al., 2021](#); [Meucci and Morana, 2025](#)). This study also reveals significant serial mediation (M3), consistent with deployment governance being associated with trustworthiness through interconnected pathways across testing, training and design stages. Collectively, these findings

provide empirical support to the cyclical (Figure 1) and interdependent structure of the CERAI lifecycle, showing that trustworthiness emerges from cumulative governance practices across all stages rather than isolated controls. A plausible organizational explanation for this pattern is that technical testing and training activities are procedurally visible but ethically opaque to internal and external stakeholders: passing a robustness or fairness test signals technical competence but does not, by itself, communicate the value commitments that were embedded earlier in the system's design. Cognitively, users and reviewers may treat design-stage ethical framing as a heuristic cue for trustworthiness (a signal of institutional intent), while testing outcomes are more often interpreted as a compliance checkbox rather than as evidence of an organization's underlying ethical commitments; when testing is decoupled from a visible ethical design rationale, it may therefore fail to move perceived trustworthiness on its own. We emphasize that this explanation is offered as a plausible post-hoc interpretation of the pattern of results rather than as an empirically tested mechanism. Constructs such as ethical opacity, stakeholder visibility and heuristic processing were not directly measured in this study, and future research incorporating these constructs is needed to test this explanation directly.

A complementary explanation, grounded in perceptual proximity, helps clarify why deployment and usage governance (H3) exhibits the strongest direct relationship with perceived trustworthiness: deployment-stage mechanisms such as monitoring, audit trails and user-facing accountability structures are directly observable to the stakeholders whose trust is being measured, and conceptually overlap with the reliance- and confidence-based content of the AI trustworthiness construct itself. Design-stage ethics and testing-stage activities, by contrast, are enacted upstream and are rarely directly observed by end users, so their influence on perceived trustworthiness is necessarily transmitted through more visible downstream governance rather than perceived directly. Framed this way, the differentiated pattern of direct and indirect effects across H1-H3 is better understood as differentiated perceived salience of governance activity across the AI lifecycle, rather than as definitive validation that governance effects cumulate uniformly from design through deployment.

Another particularly important insight from this study is that governance stages are sequentially interdependent rather than independent or additive. The nonsignificant direct effect of testing and training, combined with significant mediation effects, supports a mediated, stage-sequential model of governance effectiveness. This key finding extends prior work that identified testing and training as a critical but understudied stage, by showing that this stage's association with trustworthiness is transmitted through upstream design governance rather than operating independently (Strauss *et al.*, 2025). This reconciles the nonsignificant direct effect of H2 with the CERAI lifecycle theory. Rather than contradicting the theorized role of testing and training, the null direct effect clarifies that this stage's contribution to trustworthiness is theoretically and empirically indirect. The theoretical claim therefore shifts from "each lifecycle stage independently and additively builds trust" to "testing and training's value for trustworthiness is realized only when integrated with, and transmitted through, upstream design governance," refining the original stage-independent theorization rather than rejecting it.

Finally, the study contributes to theories of organizational ethical responsibility by linking governance decisions across the AI lifecycle to downstream AI trustworthiness outcomes. The findings suggest that organizations must balance foundational ethical architecture established during design with strong deployment-stage accountability mechanisms. Rather than offering a generic confirmation that ethical governance is associated with trust, this finding offers a stage-sensitive application of and empirical implication for, stakeholder theory and corporate responsibility perspectives, rather than a definitive theoretical extension (Freeman, 1984; Mitchell *et al.*, 1997; Noble, 2018; Pasquale, 2015): the differentiated governance function observed across lifecycle stages (i.e. design-stage governance operating as a largely latent mediating mechanism, and deployment-stage governance operating as the most stakeholder-visible driver of perceived trustworthiness) indicates that stakeholder salience to ethical governance is not constant across the AI lifecycle but is itself shaped by the stage at which that

governance occurs. We frame it this way because the study does not directly measure stakeholder salience dimensions such as power, legitimacy or urgency (Mitchell *et al.*, 1997), so the stakeholder-theory contribution is best read as stage-sensitive rather than definitive. Additionally, the study also helps explain why testing and training have remained understudied in prior research, as their effects operate indirectly through other lifecycle stages and are therefore difficult to detect when lifecycle stages are examined independently. In summary, collectively, these empirical findings provide an initial empirical assessment of a multidimensional organizational governance framework for AI trustworthiness.

Beyond stakeholder theory for ethical and responsible AI governance, this study also specifies a concrete empirical extension of the socio-technical perspective. Rather than treating organizational structure and technical system behavior as separately theorized influences on AI outcomes, the governance-sequencing finding shows one specific mechanism by which an organizational practice (i.e. design-stage governance) becomes embedded in, and detectable through, a technical system's downstream behavior and stakeholder perceptions of it. Finally, this contribution should be read as pertaining specifically to perceived trustworthiness, i.e. stakeholders' self-reported confidence in and reliance on AI systems, rather than to objective, independently verified trustworthy performance (e.g. measured robustness, audited fairness metrics or verified safety outcomes). This distinction between the two is central to interpreting these findings, since governance activity that shapes perceived trustworthiness need not correspond one-to-one with governance activity that improves objective system performance, and this study's findings relate only to the former.

5.2 Practical implication

The study offers valuable implications for AI governance practitioners, organizational leaders, policymakers and AI developers. First, the strong direct effect of CERAI deployment and usage on AI trustworthiness is consistent with operational-phase governance being the most immediately visible governance layer from a trustworthiness perspective. The results are consistent with deployment-phase mechanisms being associated with stakeholder trust. Organizations seeking to enhance stakeholder trust in AI systems may accordingly wish to prioritize investment in deployment-phase mechanisms that include real-time monitoring, transparent algorithmic accountability, accessible redress processes, usage audits and governance policies governing AI decision-making contexts. These are not merely compliance requirements. In this study, they are associated with higher perceived trustworthiness, though this association should not be read as proof that any single intervention causally increases trust.

Second, the statistically significant effect of CERAI design and development, combined with its role as a mediating mechanism (M1), carries an important governance sequencing implication, i.e. the findings are consistent with treating design-phase ethics and corporate responsibility as foundational to downstream governance effectiveness. The cross-sectional design does not, however, establish design governance as a strict structural prerequisite, and this recommendation should be read as a priority suggested by the observed associations rather than as a demonstrated causal requirement. Organizations that invest in deployment-stage accountability without establishing robust ethical architecture at the design stage may find their governance investments yielding diminished trust returns. This is consistent with the governance trap identified in M2. AI testing and training alone, without design foundations, does not show a significant independent association with trustworthiness. Practitioners should prioritize ethical design reviews, diverse stakeholder inclusion in system specification and governance protocols at the earliest stages of AI deployment and usage.

Concrete lifecycle-spanning oversight tools that leaders can adopt immediately include: (1) a stage-tagged accountability dashboard that logs which CERAI governance controls (design review sign-off, bias/robustness test results, deployment monitoring alerts) were completed at each lifecycle stage and by whom; (2) continuous algorithmic audit frameworks that re-run

fairness and robustness tests on production data on a recurring schedule rather than only pre-launch, with automatic escalation when drift is detected and (3) a design-to-deployment traceability log that links each deployment-stage governance action back to the design-stage ethical requirement it operationalizes, so that design-stage ethical architecture can be verified as a precondition before testing or deployment sign-off.

Third, the serial mediation pathway (M3) provides practitioners with actionable evidence for structuring AI governance investment across the lifecycle. The empirically supported sequence of design and development ethics being associated with testing and training governance, which in turn is associated with deployment and usage trustworthiness, suggests that CERAI governance resources should not be concentrated at a single stage. Governance program managers and AI managers should design integrated, *stage-spanning oversight frameworks* that ensure ethical conditions are built, tested and maintained across all three lifecycle stages. This shifts governance from event-based interventions (e.g. ethical reviews before launch) toward continuous, lifecycle-integrated processes. Fourth, the nonsignificance of the direct effect in H2 carries a cautionary message for organizations that treat AI testing as a primarily technical exercise. Such an approach, focused on accuracy, performance benchmarks and bias metrics, risks failure unless embedded within a broader ethical and responsible governance framework established at the design stage. Indeed, technical testing divorced from ethical architecture shows no significant independent association with trustworthiness outcomes in this study. Practitioners should reframe testing and training governance as ethically implanted activities, with explicit connections to the normative frameworks established in design and with documented accountability for how testing choices reflect organizational ethical commitments (Meucci and Morana, 2025).

Fifth, for AI policy and regulatory bodies, the findings support governance frameworks that extend oversight across all lifecycle stages rather than concentrating regulatory requirements at deployment. The evidence that upstream design governance mediates the trustworthiness effects of downstream stages suggests that regulatory instruments that focus solely on deployment accountability (such as post-deployment audits or output-focused compliance checks) are insufficient. Lifecycle-spanning regulatory frameworks that mandate ethical governance standards at the design, testing and deployment stages are consistent with this study's results, though a single cross-sectional survey of US AI professionals cannot by itself establish the case for specific regulatory mandates.

Finally, the results contribute to the growing effort to move AI governance from normative aspiration to evidence-based practice (Hagendorff, 2020; Leon, 2026). By empirically testing a CERAI lifecycle framework using primary data from AI experts, this study provides evidence that ethical governance can be measured, tested and improved, and not merely declared. This provides meaningful and practical implications for organizations developing AI accountability dashboards, governance maturity models and internal audit frameworks that can anchor organizational assessment and improvement processes. These practical recommendations should also be read in light of organizational context. The appropriate weighting of design, testing and deployment-stage governance investment likely varies across industry sectors, the risk level of the specific AI application and whether an organization develops its AI systems in-house, procures them from third-party vendors or primarily uses pre-built AI tools. None of these were modeled in the present study. Practitioners operating in high-risk domains or relying on third-party AI systems, in particular, may need to adapt the recommended emphasis on upstream design governance to the degree of direct control they hold over that stage of the AI lifecycle.

6. Conclusion

This study develops a lifecycle-based framework for CERAI and provides preliminary empirical evidence based on survey data collected from AI professionals in the United States.

The study provides an initial empirical foundation for the CERAI framework while identifying important opportunities for practitioners and researchers.

There are three key implications for practitioners. First, because design-stage governance was found to mediate the trustworthiness effects of later lifecycle stages, organizations should establish substantive ethical review checkpoints, including beneficence, nonmaleficence, autonomy and justice criteria, during design and development before committing resources to downstream testing and deployment safeguards. Second, because deployment-and-usage governance showed the strongest direct association with perceived trustworthiness, organizations should prioritize stakeholder-visible accountability mechanisms, including audit trails, user-facing transparency disclosures and designated oversight roles. These mechanisms can provide visible evidence of organizational commitment to trustworthy AI. Third, because testing and training governance contributed to trustworthiness indirectly through design-stage ethics, practitioners responsible for AI testing and validation should explicitly document how testing protocols operationalize the ethical commitments established during design. Testing should therefore be treated not simply as a technical compliance activity but as part of an integrated governance process.

Several limitations qualify the interpretation and generalizability of these findings. First, although nonresponse bias checks based on response timing, firm size, job position and AI leadership revealed no significant differences, the study could not assess nonresponse bias related to industry-specific governance maturity because of data availability. Generalizability across this dimension should therefore be interpreted with caution. Second, the cross-sectional, single-source survey design prevents the study from establishing causal or temporal relationships. Thus, the serial mediation findings should be interpreted as associations consistent with the hypothesized governance-sequencing logic rather than as evidence that governance at one lifecycle stage causes outcomes at another. Third, although the sample provided sufficient statistical power, participants were AI professionals recruited at an AI-focused conference. A larger and more heterogeneous sample drawn from multiple AI-related professional networks, events and organizational contexts could have strengthened the generalizability of the findings. Fourth, the survey did not include an explicit screening question confirming respondents' direct responsibility for or involvement in AI governance. Finally, the study's US-based sample limits the extent to which the findings can be generalized across different regulatory, institutional and cultural environments. The findings should therefore be viewed as preliminary empirical evidence supporting the CERAI lifecycle perspective rather than as evidence of universal governance relationships.

These limitations also point to several opportunities for future research. First, longitudinal and multi-source studies should examine whether the governance-sequencing relationships identified here reflect causal and temporal processes. Such research could also determine whether investments in lifecycle governance produce sustained or compounding effects on AI trustworthiness as AI systems evolve. Second, future studies should directly measure the mechanisms proposed to explain the observed relationships, including ethical opacity, stakeholder visibility and stakeholder salience dimensions such as power, legitimacy and urgency. This would move the research beyond post-hoc interpretation toward direct empirical testing of the underlying mechanisms. Third, researchers should examine boundary conditions and moderators, including organizational size, industry sector, AI maturity and the presence of dedicated AI ethics functions. Replication across organizations that develop AI in-house versus those that procure or use third-party AI systems could further clarify when and how lifecycle governance influences trustworthiness. Finally, replication across diverse regulatory, institutional and cultural contexts, including the Global South, East Asia, the European Union and emerging economies, would help establish the external validity of the CERAI framework. Together, these research directions can strengthen the empirical foundation of CERAI and advance understanding of how ethical and responsible AI governance can be effectively institutionalized across the AI lifecycle.

Disclaimer

AI-enabled word processor was used to improve language.

Appendix**Survey instrument****(1) CERAi design and development***Nonmaleficence by design*

DD-NMF1 should be designed to prevent harmful outcomes.

DD-NMF2 risks should be evaluated regularly.

DD-NMF3 design should have safeguards to prevent misuse.

DD-NMF4 design should reduce potential harm.

Beneficence by design

DD-BEF1 should be designed to improve user's decision quality.

DD-BEF2 should be designed to create positive outcomes for users.

DD-BEF3 should be designed to improve outcome effectiveness.

DD-BEF4 should be designed to enhance human capabilities.

DD-BEF5 should be designed to improve various stakeholder and societal outcomes.

Autonomy by design

DD-AMY1 should be designed for users to override AI decisions when necessary.

DD-AMY2 should be designed to support rather than replace human decision-making.

DD-AMY3 should be designed for users to retain control over how AI recommendations are applied.

DD-AMY4 should be designed for users' oversight.

Justice (fairness) by design

DD-FNS1 should be designed to prevent potential bias.

DD-FNS2 should be designed for fairness in AI outcomes.

DD-FNS3 should be designed to avoid discrimination.

DD-FNS4 should be designed not to disadvantage specific social groups.

(2) CERAi testing and training*Robustness*

TG-RBS1 should be regularly tested and trained for reliability.

TG-RBS2 should be regularly tested and trained for accuracy.

TG-RBS3 should be tested and trained extensively before deployment.

TG-RBS4 should be regularly tested and trained for consistency.

TG-RBS5 should be regularly tested and trained for resiliency.

Fairness

TG-FNS1 Data used to train and test AI models should be checked for bias.

TG-FNS2 AI outcomes should be monitored for fairness across groups during AI model training and testing.

TG-FNS3 Algorithmic bias should be mitigated during AI model training and testing.

TG-FNS4 AI models should be evaluated for discriminatory impacts during AI model training and testing.

(3) CERAi deployment and usage (user experience)

Explainability

UX-EXP1 AI decision processes should provide clear explanations.

UX-EXP2 Users need to be knowledgeable about how AI systems produce results.

UX-EXP3 AI outputs should include explanations that users can interpret.

UX-EXP4 AI systems should present decision explanations in an accessible format.

UX-EXP5 Documentation should exist explaining AI model behavior.

Transparency

UX-TRP1 The organization should be transparent about how AI is used.

UX-TRP2 AI system functionality should be documented and accessible.

UX-TRP3 The organization should clearly disclose how AI systems use data.

UX-TRP4 Users should be provided with a clear understanding of the role of AI in decision-making.

UX-TRP5 AI governance policies should be clearly communicated.

Accountability

UX-ACT1 Clear responsibility should exist for decisions made by AI systems.

UX-ACT2 AI outcomes should be audited regularly.

UX-ACT3 Policies to define accountability should exist for AI usage.

UX-ACT4 Supervisors should be responsible for monitoring AI decision processes.

User privacy

UX-DPR1 Personal data used during AI usage should be securely protected.

UX-DPR2 The organization should have clear and strict data privacy policies for users.

UX-DPR3 Access to sensitive data in AI usage should be carefully controlled.

UX-DPR4 Employees should be properly trained on responsible data handling.

(4) AI trustworthiness

AI-TRS1 AI system task performance should be reliable.

AI-TRS2 AI system results should be dependable.

AI-TRS3 AI systems that follow ethical standards enhances user's confident in AI usage.

AI-TRS4 AI systems that follow ethical standards enhances user's trust in AI usage.

References

- Abraham, R., Schneider, J. and vom Brocke, J. (2019), "Data governance: a conceptual framework, structured review, and research agenda", *International Journal of Information Management*, Vol. 49, pp. 424-438, doi: [10.1016/j.ijinfomgt.2019.07.008](https://doi.org/10.1016/j.ijinfomgt.2019.07.008).
- Afifa, M.A., Nguyen, N.M. and Bui, D.V. (2025), "Environmental, social and governance (ESG) disclosure quality in developing countries: evidence from the ASEAN region", *Corporate Governance*, Vol. 25 No. 7, pp. 1728-1754, doi: [10.1108/CG-02-2024-0102](https://doi.org/10.1108/CG-02-2024-0102).

- Barocas, S. and Selbst, A.D. (2016), "Big data's disparate impact", *California Law Review*, Vol. 104 No. 3, pp. 671-732.
- Barocas, S., Hardt, M. and Narayanan, A. (2023), *Fairness and Machine Learning: Limitations and Opportunities*, fairmlbook.org.
- Binns, R. (2018), "Fairness in machine learning: lessons from political philosophy", *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, ACM, pp. 149-159.
- Castelluccia, C. and Le Métayer, D. (2019), *Understanding Algorithmic decision-making: Opportunities and Challenges*, European Parliament, available at: [https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU\(2019\)624261](https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624261)
- Cath, C. (2018), "Governing artificial intelligence: ethical, legal and technical opportunities and challenges", *Philosophical Transactions of the Royal Society A*, Vol. 376 No. 2133, 20180080, doi: [10.1098/rsta.2018.0080](https://doi.org/10.1098/rsta.2018.0080).
- Chin, W.W. (1998), "The partial least squares approach to structural equation modeling", in Marcoulides, G.A. (Ed.), *Modern Methods for Business Research*, Lawrence Erlbaum Associates, pp. 295-336.
- Chiodo, M., Müller, D. and Sienknecht, M. (2024), "Educating AI developers to prevent harmful path dependency in AI resort-to-force decision making", *Australian Journal of International Affairs*, Vol. 78 No. 2, pp. 210-219, doi: [10.1080/10357718.2024.2327366](https://doi.org/10.1080/10357718.2024.2327366).
- Cihon, P., Kleinaltenkamp, M.J., Schuett, J. and Baum, S.D. (2021), "AI certification: advancing ethical practice by reducing information asymmetries", *IEEE Transactions on Technology and Society*, Vol. 2 No. 4, pp. 200-209, doi: [10.1109/TTS.2021.3077595](https://doi.org/10.1109/TTS.2021.3077595).
- Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates.
- Cubric, M. (2020), "Drivers, barriers and social considerations for AI adoption in business and management: a tertiary study", *Technology in Society*, Vol. 62, 101257, doi: [10.1016/j.techsoc.2020.101257](https://doi.org/10.1016/j.techsoc.2020.101257).
- de Fine Licht, J. and de Fine Licht, K. (2020), "Artificial intelligence, transparency, and public decision-making", *AI & Society*, Vol. 35 No. 4, pp. 917-926, doi: [10.1007/s00146-020-00960-w](https://doi.org/10.1007/s00146-020-00960-w).
- de Haes, S. and Van Grembergen, W. (2009), "An exploratory study into IT governance implementations and its impact on business/IT alignment", *Information Systems Management*, Vol. 26 No. 2, pp. 123-137, doi: [10.1080/10580530902794786](https://doi.org/10.1080/10580530902794786).
- de Lucas López, A.J., Luengo Vera, C. and Álvarez Miguel, Á.J. (2026), "Artificial intelligence as a catalyst for the circular economy: a mixed-methods analysis of sustainable production ecosystems (2023-2024)", *The International Entrepreneurship and Management Journal*, Vol. 22 No. 2, 59, doi: [10.1007/s11365-026-01176-y](https://doi.org/10.1007/s11365-026-01176-y).
- DiMaggio, P.J. and Powell, W.W. (1983), "The iron cage revisited: institutional isomorphism and collective rationality in organizational fields", *American Sociological Review*, Vol. 48 No. 2, pp. 147-160, doi: [10.2307/2095101](https://doi.org/10.2307/2095101).
- Dong, X.L. and Srivastava, D. (2015), *Big Data Integration*, Morgan & Claypool.
- Doty, D.H. and Glick, W.H. (1998), "Common methods bias: does common methods variance really bias results?", *Organizational Research Methods*, Vol. 1 No. 4, pp. 374-406, doi: [10.1177/109442819814002](https://doi.org/10.1177/109442819814002).
- Edwards, J. (2026), "Evaluating AI robustness and response strategies", in Edwards, J. (Ed.), *Adversarial Machine Learning*, Wiley. doi: [10.1002/9781394402069.ch14](https://doi.org/10.1002/9781394402069.ch14).
- European Parliament (2024), "Regulation (EU) 2024/1689 of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act)", *Official Journal of the European Union*, (accessed 2026).
- Floridi, L. and Cows, J. (2019), "A unified framework of five principles for AI in society", *Harvard Data Science Review*, Vol. 1 No. 1, doi: [10.1162/99608f92.8cd550d1](https://doi.org/10.1162/99608f92.8cd550d1).
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018), "AI4People — An ethical

- framework for a good AI society: opportunities, risks, principles, and recommendations”, *Minds and Machines*, Vol. 28 No. 4, pp. 689-707, doi: [10.1007/s11023-018-9482-5](https://doi.org/10.1007/s11023-018-9482-5).
- Flynn, B.B., Schroeder, R.G. and Sakakibara, S. (1994), “A framework for quality management research and an associated measurement instrument”, *Journal of Operations Management*, Vol. 11 No. 4, pp. 339-366, doi: [10.1016/s0272-6963\(97\)90004-8](https://doi.org/10.1016/s0272-6963(97)90004-8).
- Fornell, C. and Larcker, D.F. (1981), “Evaluating structural equation models with unobservable variables and measurement error”, *Journal of Marketing Research*, Vol. 18 No. 1, pp. 39-50, doi: [10.1177/002224378101800104](https://doi.org/10.1177/002224378101800104).
- Freeman, R.E. (1984), *Strategic Management: a Stakeholder Approach*, Pitman.
- Friedman, B., Kahn, P.H. Jr and Borning, A. (2002), *Value Sensitive Design: Theory and Methods*, [Technical report], University of Washington.
- Frohlich, M.T. (2002), “Techniques for improving response rates in OM survey research”, *Journal of Operations Management*, Vol. 20 No. 1, pp. 53-62, doi: [10.1016/s0272-6963\(02\)00003-7](https://doi.org/10.1016/s0272-6963(02)00003-7).
- Ganesh, N., Siddineni, D., Reddy, V.V., Lateef, K. and Sharma, R. (2025), “Corporate governance in the age of AI: ethical oversight and accountability frameworks”, *Journal of Information Systems Engineering and Management*, Vol. 10, pp. 1141-1148, doi: [10.52783/jeism](https://doi.org/10.52783/jeism).
- George, D. and Mallery, P. (2010), *SPSS for Windows Step by Step: a Simple Guide and Reference*, 17.0 Update, 10th ed., Allyn & Bacon.
- Gupta, S., Masmoudi, M., Gupta, N. and Srivastava, R. (2026), “Synergizing Society 5.0 and SDGs: a mixed-method roadmap for sustainable manufacturing transformation”, *Business Strategy and Development*, Vol. 9 No. 1, e70316, doi: [10.1002/bsd2.70316](https://doi.org/10.1002/bsd2.70316).
- Hagendorff, T. (2020), “The ethics of AI ethics: an evaluation of guidelines”, *Minds and Machines*, Vol. 30 No. 1, pp. 99-120, doi: [10.1007/s11023-020-09517-8](https://doi.org/10.1007/s11023-020-09517-8).
- Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1995), *Multivariate Data Analysis with Readings*, 4th ed., Prentice Hall.
- Hansson, S.O. (2017), *The Ethics of Technology: Methods and Approaches*, Bloomsbury.
- Henseler, J., Ringle, C.M. and Sarstedt, M. (2015), “A new criterion for assessing discriminant validity in variance-based structural equation modeling”, *Journal of the Academy of Marketing Science*, Vol. 43 No. 1, pp. 115-135, doi: [10.1007/s11747-014-0403-8](https://doi.org/10.1007/s11747-014-0403-8).
- Höllerer, M.A. (2013), “From taken-for-granted to explicit commitment: the rise of CSR in a corporatist country”, *Journal of Management Studies*, Vol. 50 No. 4, pp. 573-606, doi: [10.1111/joms.12029](https://doi.org/10.1111/joms.12029).
- Jobin, A., Ienca, M. and Vayena, E. (2019), “The global landscape of AI ethics guidelines”, *Nature Machine Intelligence*, Vol. 1 No. 9, pp. 389-399, doi: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2).
- Kaplan, A. and Haenlein, M. (2019), “Siri, Siri, in my hand, who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence”, *Business Horizons*, Vol. 62 No. 1, pp. 15-25, doi: [10.1016/j.bushor.2018.08.004](https://doi.org/10.1016/j.bushor.2018.08.004).
- Khatri, V. and Brown, C.V. (2010), “Designing data governance”, *Communications of the ACM*, Vol. 53 No. 1, pp. 148-152, doi: [10.1145/1629175.1629210](https://doi.org/10.1145/1629175.1629210).
- Kock, N. (2017), “Common method bias: a full collinearity assessment method for PLS-SEM”, in Latan, H. and Noonan, R. (Eds), *Partial Least Squares Path Modeling*, Springer, pp. 245-257.
- Latour, B. (2005), *Reassembling the Social: an Introduction to actor-network Theory*, Oxford University Press.
- Lee, F. (2025), “Reassembling agency: epistemic practices in the age of artificial intelligence”, *Sociologisk Forskning*, Vol. 62 Nos 1-2, pp. 43-58, doi: [10.37062/sf.62.27824](https://doi.org/10.37062/sf.62.27824).
- Leon, M. (2026), “Lifecycle-based governance to build reliable ethical AI systems”, *Systems Research and Behavioral Science*, Vol. 43 No. 3, pp. 1116-1132, doi: [10.1002/sres.70014](https://doi.org/10.1002/sres.70014).
- Liapis, C.M., Fazakis, N., Kotsiantis, S. and Dimakopoulos, Y. (2026), “Ethics in artificial intelligence: a cross-sectoral review of 2019-2025”, *Informatics*, Vol. 13 No. 4, 51, doi: [10.3390/informatics13040051](https://doi.org/10.3390/informatics13040051).

- Maria, E., Cuevas, R. and Baquerio, R.C. (2026), *The Seven Parameters of Ethical, Legal, and Social Responsibility Structures to Protect the Human Workforce in Business Management in the Age of Artificial Intelligence*, ResearchGate. doi: [10.13140/RG.2.2.15218.16327](https://doi.org/10.13140/RG.2.2.15218.16327).
- Mayer, A., van den Broek, E. and Karacic, T. (2026), "When employees are held accountable for AI-generated decisions", *Harvard Business Review*, pp. 1-10.
- McKnight, D.H., Carter, M., Thatcher, J.B. and Clay, P.F. (2011), "Trust in a specific technology: an investigation of its components and measures", *ACM Transactions on Management Information Systems*, Vol. 2 No. 2, pp. 12-25, doi: [10.1145/1985347.1985353](https://doi.org/10.1145/1985347.1985353).
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021), "A survey on bias and fairness in machine learning", *ACM Computing Surveys*, Vol. 54 No. 6, pp. 115-135, doi: [10.1145/3457607](https://doi.org/10.1145/3457607).
- Metcalf, J., Moss, E. and Boyd, D. (2019), "Owning ethics: corporate logics, Silicon valley, and the institutionalization of AI ethics", *Social Research: International Quarterly*, Vol. 86 No. 2, pp. 449-476, doi: [10.1353/sor.2019.0022](https://doi.org/10.1353/sor.2019.0022).
- Meucci, M. and Morana, M. (2025), *OWASPAI Testing Guide*, OWASP, available at: <https://owasp.org/www-project-ai-testing-guide/>
- Meystel, A., Albus, J., Messina, E. and Leedom, D. (2003), *Performance Measures for Intelligent Systems: Measures of Technology Readiness*, National Institute of Standards and Technology.
- Mitchell, R.K., Agle, B.R. and Wood, D.J. (1997), "Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts", *Academy of Management Review*, Vol. 22 No. 4, pp. 853-886, doi: [10.2307/259247](https://doi.org/10.2307/259247).
- Mittelstadt, B. (2019), "Principles alone cannot guarantee ethical AI", *Nature Machine Intelligence*, Vol. 1 No. 11, pp. 501-507, doi: [10.1038/s42256-019-0114-4](https://doi.org/10.1038/s42256-019-0114-4).
- Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016), "The ethics of algorithms: mapping the debate", *Big Data and Society*, Vol. 3 No. 2, pp. 1-21, doi: [10.1177/2053951716679679](https://doi.org/10.1177/2053951716679679).
- Mittelstadt, B., Russell, C. and Wachter, S. (2019), "Explaining explanations in AI", *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, ACM, pp. 279-288.
- Morley, J., Machado, C.C.V., Burr, C., Cowls, J., Joshi, I., Taddeo, M. and Floridi, L. (2020), "The ethics of AI in health care: a mapping review", *Social Science and Medicine*, Vol. 260, 113172, doi: [10.1016/j.socscimed.2020.113172](https://doi.org/10.1016/j.socscimed.2020.113172).
- Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J. and Floridi, L. (2021), "Ethics as a service: a pragmatic operationalisation of AI ethics", *Minds and Machines*, Vol. 31 No. 2, pp. 239-256, doi: [10.1007/s11023-021-09563-w](https://doi.org/10.1007/s11023-021-09563-w).
- Nastoska, A., Jancheska, B., Rizinski, M. and Trajanov, D. (2025), "Evaluating trustworthiness in AI: risks, metrics, and applications across industries", *Electronics*, Vol. 14 No. 13, 2717, doi: [10.3390/electronics14132717](https://doi.org/10.3390/electronics14132717).
- National Institute of Standards and Technology (2023), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, U.S. Department of Commerce, available at: <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>
- Nguyen, T.T., Nguyen, N.M. and Bui, D.V. (2026), "Unlocking sustainable value through intelligent information systems: the interplay of AI, governance, and reporting quality in ASEAN", *VINE Journal of Information and Knowledge Management Systems*, pp. 1-23, doi: [10.1108/VJKMS-09-2025-0409](https://doi.org/10.1108/VJKMS-09-2025-0409).
- Noble, S.U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press.
- Pahune, S., Akhtar, Z., Mandapati, V. and Siddique, K. (2025), "The importance of AI data governance in large language models", *Big Data and Cognitive Computing*, Vol. 9 No. 6, 147, doi: [10.3390/bdcc9060147](https://doi.org/10.3390/bdcc9060147).
- Pan, Y., Huang, Z., Lev, B., Xu, L. and Olson, D.L. (2026), "A literature review on the artificial intelligence in manufacturing systems under Industry 5.0", *International Journal of Production Research*, Vol. 64 No. 12, pp. 5120-5161, doi: [10.1080/00207543.2026.2623537](https://doi.org/10.1080/00207543.2026.2623537).

- Papagiannidis, E., Mikalef, P. and Conboy, K. (2025), "Responsible artificial intelligence governance: a review and research framework", *The Journal of Strategic Information Systems*, Vol. 34 No. 2, 101885, doi: [10.1016/j.jsis.2024.101885](https://doi.org/10.1016/j.jsis.2024.101885).
- Pariser, E. (2011), *The Filter Bubble: what the Internet is Hiding from you*, Penguin Press.
- Pasquale, F. (2015), *The Black Box Society: the Secret algorithms that Control Money and Information*, Harvard University Press.
- Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y. and Podsakoff, N.P. (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88 No. 5, pp. 879-903, doi: [10.1037/0021-9010.88.5.879](https://doi.org/10.1037/0021-9010.88.5.879).
- Rai, A. (2020), "Explainable AI: from Black box to glass box", *Journal of the Academy of Marketing Science*, Vol. 48 No. 1, pp. 137-141, doi: [10.1007/s11747-019-00710-5](https://doi.org/10.1007/s11747-019-00710-5).
- Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020), "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing", *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, ACM, pp. 33-44.
- Rathkopf, C. (2025), "Hallucination, reliability, and the role of generative AI in science", [Working paper], arXiv, available at: <https://arxiv.org/html/2504.08526v1>
- Reynolds, G.W. (2022), *Ethics in Information Technology*, 7th ed., Cengage Learning.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016), "'Why should I trust you?': explaining the predictions of any classifier", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1135-1144.
- Samek, W., Wiegand, T. and Müller, K.-R. (2017), "Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models", [Preprint], arXiv, available at: <https://arxiv.org/abs/1708.08296>
- Schiff, D., Biddle, J., Borenstein, J. and Laas, K. (2020), "What's next for AI ethics, policy, and governance? A global overview", *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, pp. 153-158.
- Shan, F., Mao, S., Lu, Y. and Li, S. (2024), "Differential privacy federated learning: a comprehensive review", *International Journal of Advanced Computer Science and Applications*, Vol. 15 No. 7, doi: [10.14569/IJACSA.2024.0150722](https://doi.org/10.14569/IJACSA.2024.0150722).
- Shmueli, G., Sarstedt, M., Hair, J.F., Cheah, J.-H., Ting, H., Vaithilingam, S. and Ringle, C.M. (2019), "Predictive model assessment in PLS-SEM: guidelines for using PLSpredict", *European Journal of Marketing*, Vol. 53 No. 11, pp. 2322-2347, doi: [10.1108/EJM-02-2019-0189](https://doi.org/10.1108/EJM-02-2019-0189).
- Singh, P. and Singh, V. (2024), "The power of AI: enhancing customer loyalty through satisfaction and efficiency", *Cogent Business and Management*, Vol. 11 No. 1, 2326107, doi: [10.1080/23311975.2024.2326107](https://doi.org/10.1080/23311975.2024.2326107).
- Strauss, I., Moure, I., O'Reilly, T. and Rosenblat, S. (2025), "Real-world gaps in AI governance and research: AI safety and reliability in everyday deployments", [Preprint], arXiv, available at: <https://arxiv.org/html/2505.00174v2>
- Tandoc, E.C. Jr, Lim, Z.W. and Ling, R. (2018), "Defining 'fake news': a typology of scholarly definitions", *Digital Journalism*, Vol. 6 No. 2, pp. 137-153, doi: [10.1080/21670811.2017.1360143](https://doi.org/10.1080/21670811.2017.1360143).
- Teece, D.J., Pisano, G. and Shuen, A. (1997), "Dynamic capabilities and strategic management", *Strategic Management Journal*, Vol. 18 No. 7, pp. 509-533.
- Trist, E.L. and Bamforth, K.W. (1951), "Some social and psychological consequences of the longwall method of coal-getting", *Human Relations*, Vol. 4 No. 1, pp. 3-38, doi: [10.1177/001872675100400101](https://doi.org/10.1177/001872675100400101).
- Uren, V. and Edwards, J.S. (2023), "Technology readiness and the organizational journey towards AI adoption: an empirical study", *International Journal of Information Management*, Vol. 68, 102594, doi: [10.1016/j.ijinfomgt.2022.102588](https://doi.org/10.1016/j.ijinfomgt.2022.102588).

- Wachter, S., Mittelstadt, B. and Russell, C. (2018), "Counterfactual explanations without opening the black box: automated decisions and the GDPR", *Harvard Journal of Law and Technology*, Vol. 31 No. 2, pp. 841-887.
- Werthner, H., Prem, E., Lee, E.A. and Ghezzi, C. (2023), *Perspectives on Digital Humanism*, Springer.
- Yahoo Tech (2024), "This Claude-powered AI agent deleted a company's whole database — and then gloated about it", Yahoo Tech, available at: <https://www.yahoo.com/tech/ai/article/this-claude-powered-ai-agent-deleted-a-companys-whole-database-and-then-gloated-about-it-165838948.html>
- Yu, L. and Li, Y. (2022), "Artificial intelligence decision-making transparency and employees' trust: the parallel multiple mediating effect of effectiveness and discomfort", *Behavioral Sciences*, Vol. 12 No. 5, 127, doi: [10.3390/bs12050127](https://doi.org/10.3390/bs12050127).
- Zhao, X., Lynch, J.G. Jr and Chen, Q. (2010), "Reconsidering Baron and Kenny: myths and truths about mediation analysis", *Journal of Consumer Research*, Vol. 37 No. 2, pp. 197-206, doi: [10.1086/651257](https://doi.org/10.1086/651257).
- Zou, H., Chan, K.I., Pang, P.C.-I., Manditereza, B. and Shih, Y.-H. (2025), "Factors influencing the reported intention of higher vocational computer science students in China to use AI after ethical training: a study in Guangdong Province", *Education Sciences*, Vol. 15 No. 11, 1431, doi: [10.3390/educsci15111431](https://doi.org/10.3390/educsci15111431).
- Zou, H., Chan, K.I., Pang, P., Manditereza, B. and Shih, Y.-H. (2026), "To use but not to depend: pedagogical novelty and the cognitive brake of ethical awareness in computer science students' adoption of generative AI", *Education Sciences*, Vol. 16 No. 2, 311, doi: [10.3390/educsci16020311](https://doi.org/10.3390/educsci16020311).
- Zuur, A.F., Ieno, E.N. and Elphick, C.S. (2010), "A protocol for data exploration to avoid common statistical problems", *Methods in Ecology and Evolution*, Vol. 1 No. 1, pp. 3-14, doi: [10.1111/j.2041-210x.2009.00001.x](https://doi.org/10.1111/j.2041-210x.2009.00001.x).

Corresponding author

Rupak Rauniar can be contacted at: rauniarr@uhd.edu