

DOI: https://doi.org/10.48009/1_iis_2026_127

Artificial Intelligence ethical frameworks within organizations: Empirical validation of an instrument

Carol Springer Sargent, *Mercer University*, sargent_cs@mercerv.edu

Alex Koohang, *Middle Georgia University*, alex.koohang@mga.edu

Marc Miller, *Middle Georgia University*, marc.miller@mga.edu

Tara Underwood, *Middle Georgia University*, tara.underwood@mga.edu

Kevin Floyd, *Middle Georgia University*, kevin.floyd@mga.edu

Tabitha Floyd, *Middle Georgia University*, tabitha.floyd@mga.edu

Abstract

This study empirically validates an instrument for assessing users' concerns about ethical frameworks for Artificial Intelligence (AI) within organizations. The instrument was originally conceptualized with 12 constructs: data responsibility, accountability, data privacy, fairness, explainability, transparency, robustness, moral agency, value alignment, technology misuse, leadership, and trust. The research examines the reliability and interpretability of these constructs relative to their associated variables. Data from 244 subjects across various organizations in the USA were analyzed using Exploratory Factor Analysis (EFA) in SPSS (Version 28). After removing items with significant cross-loadings, the final factor solution was established. The EFA revealed that although 12 theoretical constructs were proposed, users perceive AI ethical concerns through a consolidated 8-factor structure. Key mergers occurred between Data Responsibility and Accountability (Component 1), Robustness and Technology Misuse (Component 2), and Explainability and Transparency (Component 4). Conversely, Data Privacy, Fairness, Leadership, and Trust emerged as distinct, standalone pillars with high internal consistency. The results suggest a high degree of conceptual overlap among stakeholders, with technical implementation and organizational oversight viewed as integrated dimensions. This study offers a streamlined, validated roadmap for organizations to evaluate their ethical standing. It emphasizes that ethical AI adoption requires more than technical robustness; it also needs a cultural shift focused on leadership, strategic vision, and trust.

Keywords: AI ethics frameworks, data responsibility, accountability, data privacy, fairness, explainability, transparency, robustness, moral agency, value alignment, technology misuse, leadership, trust

Introduction

The rapid and dynamic evolution of Artificial Intelligence (AI) across public, commercial, and personal life introduces many ethical challenges, including "black box" non-deterministic algorithmically derived outputs, massive data scraping from potentially biased or unreliable sources, and diffusion of responsibility for unintended or ill-intended consequences of AI software. For instance, is the AI designer responsible for a weak implementation (which may include tuning the model after design) and for a wide range of possible uses the designer never intended? Some organizations, professional groups, and governmental authorities have considered codes of digital ethics, AI frameworks, or relevant principles that might serve as a multi-

use tool for understanding ethical issues with AI use. Agreement is difficult because some stakeholders prioritize widening ethical responsibility, others emphasize versatility within basic principles, and some prefer more prescriptive guidelines. What is needed is a shared understanding of foundational AI ethical principles that resonate with diverse stakeholders and generate trust in and adoption of AI systems.

Based on an extensive review of the literature, this study groups the most widely-agreed-upon AI ethical issues into 12 frameworks. The purpose of this study is to empirically validate an instrument comprising these 12 constructs (data responsibility, accountability, data privacy, fairness, explainability, transparency, robustness, moral agency, value alignment, technology misuse, leadership, and trust) to assess users' concerns about AI ethical frameworks in organizations. The following research question will be addressed. Are the constructs/components (data responsibility, accountability, data privacy, fairness, explainability, transparency, robustness, moral agency, value alignment, technology misuse, leadership, and trust) reliable and interpretable in relation to their associated variables?

Literature review

While AI literature does not currently offer a generally accepted checklist of guiding principles for ethical AI design and use, there are major frameworks that are widely discussed. Stakeholders use a range of terms to describe AI ethical frameworks, making it difficult to distill the core principles these frameworks need. To distill the most widely agreed-upon AI ethics frameworks, we have identified ten, each built around a set of core principles. These frameworks are data responsibility, accountability, data privacy, fairness, explainability, transparency, robustness, moral agency, value alignment, technology misuse, leadership, and trust. In the next section, we will discuss these frameworks.

AI Ethical Framework: Data Responsibility

Data responsibility, sometimes used interchangeably with accountability (Jang et al., 2022), centers on ensuring human oversight of the AI model and clearly defining who is responsible for AI-related errors or harms (AI HLEG, 2019; de-Lima-Santos et al., 2024; Deloitte AI Institute, 2024; Lekadir et al., 2024). Stakeholders must assess models, and humans-in-the-loop must remain vigilant for potential harm (AI HLEG, 2019; Makridis et al., 2024). AI models must be monitored for potential misuse and harmful adaptations, and designers need to define the intended use and user requirements early (Lekadir et al., 2024; Pichai, 2018). While it seems reasonable that the AI developer should take responsibility for the design, what about those who purchase the AI program and deploy it in a novel way? Across many authors, the theme is that those operating the AI system, not just the designer, should take responsibility for errors or harm (Jang et al., 2022). Another approach is the multi-stakeholder responsibility model, which spreads responsibility across designers, implementors, operators, governance bodies, and leaders adopting AI, based on the degree of control over negative outcomes (Badal et al., 2023; Gong, 2026; Wecker et al., 2026).

AI Ethical Framework: Accountability

Like responsibility, accountability relies on self-governance, with AI actors providing clear, timely, and accessible notice of decisions about the AI system. Organizations committed to accountability would include multi-stakeholder participation, such as users, cross-functional teams, and independent evaluation (Lekadir et al., 2024; Stix, 2021; The White House, 2022). Furthermore, the accountable AI system would have a process for feedback, responses to feedback, and appeals for redress. This AI theme is distinct from responsibility (as in "who" is responsible) and from accountability, as the governance mechanism in place to redress harm or adverse impact caused by third parties, such as an AI ethics board (Hirsch et al., 2024; Mökander et al., 2022). The NIST suggests that accountability would include communicating incidents and

errors to AI actors and affected user communities, with the process for tracking, responding, and recovering from incidents clearly articulated (NIST, 2023).

AI Ethical Framework: Data Privacy

Data privacy, the right to decide who sees users' personal information, encompasses identity, task, and attribute privacy and affects many technologies, not just AI. Privacy issues with AI systems are mentioned in 80% of AI ethics articles, making it the most widely agreed-upon ethical issue (Hagendorff, 2019). Addressing the ethical concern about AI systems retaining users' data requires notice and consent. If consent is presumed by device privacy settings, some ask whether the default should be "opt-out" because finding and setting privacy preferences is burdensome for users (The White House, 2022). Going further, some insist that ethical handling of personal data (and synthetic data created from private data) should address de-identification and improper re-identification (AI HLEG, 2019; Cassidy et al., 2022). AI can impersonate others' voices and writing styles, further raising ethical issues regarding the use of private data. AI can also scrape massive amounts of data from historical public archives, some of which may be "leaky" and contain personally identifiable information (PI), such as driver's license data. Ensuring that AI controls and uses user data, especially private data that AI may not properly flag as requiring consent, is a major ethical concern.

AI Ethical Framework: Fairness

Fairness, like data privacy, is a concern in over 80% of AI ethics articles (Hagendorff, 2019). Fairness is a complex concept because some use it interchangeably with inclusion (Lekadir et al., 2024). Fairness and inclusion raise ethical questions about unwanted bias, how vulnerable groups may be affected, and the disparate impact of AI systems on individual or community attributes (Hirsch et al., 2024). AI models process data impartially, but if trained on historical, non-representative datasets, their recommendations may reflect or perpetuate unjust correlations present in the training data (Jang et al., 2022; Pichai, 2018; The White House, 2022). Removing historical discrimination from training data is nearly impossible (Habbal et al., 2024). AI is a correlation finder that identifies more correlations as it learns, assuming past correlations will recur, including unjust ones (Colin Clemente Jones, 2023). Detecting and addressing unfair, exclusionary, or skewed patterns in AI output is difficult due to the opaque nature of deep learning (Liang & Acuna, 2020). While insisting on training with representative datasets makes sense (Lekadir et al., 2024), AI designers may lack access to broad, diverse datasets or may not know the extent of the diversity and non-representativeness in the training materials (Bao et al., 2023; Saba & Pretorius, 2024). Complicating the ethical problems, unfairness may be noticed only after deployment (Köchling & Wehner, 2023). This is especially concerning when decision-making processes that utilize the potentially biased algorithms of AI are relevant to patient care and treatment decisions by clinicians (Giovanola & Tiribelli, 2023). While this may be a challenging area, there is broad agreement that fairness in AI model design and the use of representative datasets matter to AI stakeholders.

AI Ethical Framework: Explainability

An AI system's ability to show how its models obtain training data, how the model has been tuned after initial training, how the model's algorithm makes recommendations, and how the models are secured once deployed determines explainability, a term sometimes used interchangeably with interpretability or understandability (Jang et al., 2022; Lekadir et al., 2024; Makridis et al., 2024; Taylor & Taylor, 2021). The opposite of explainability is the black box or opacity, where the model owner cannot interpret how the model learns, debug unexpected results, or understand why a model failed (Hirsch et al., 2024). AI system designers grapple with this ethical issue because explainability depends largely on the target user and the context that guides the model's output. Despite possible challenges in making AI systems understandable, being able to interpret, trace, and explicate AI's actions is a key element in determining whether AI's

algorithms are ethically designed (Badal et al., 2023; Cassidy et al., 2022; Hirsch et al., 2024; IBM, 2024; Makridis et al., 2024; NIST, 2023).

AI Ethical Framework: Transparency

Transparency of AI systems requires sharing the process and the drivers of the model's output (Hagendorff, 2019). Often, the AI model algorithms act like black boxes, opaque in the way decision-making routines work (Christoforaki & Beyan, 2022; Lekadir et al., 2024). When the decision criteria or model sequencing is hidden from the public eye, it is unclear if ethical considerations were part of the model's design (Badal et al., 2023; Hirsch et al., 2024). Increasing transparency requires communicating clearly how the model was designed and tuned (AI HLEG, 2019). Further, users need to know the key drivers of output, including the weights and the importance of inputs (Cassidy et al., 2022). An ethical AI model would reveal how user-provided details triggered the AI responses, removing the black box syndrome of many AI models (Bao et al., 2023; Dikmen & Burns, 2022; Higgins et al., 2023). For instance, in healthcare, a provider might want to see the deterministic process of how records converged to recommend a medical finding, a struggle for a model that learns continuously (Bergquist et al., 2024). A fully transparent system may need to go beyond sharing details about the model, such as the transparency of persons (Badal et al., 2023).

AI Ethical Framework: Robustness

AI system robustness reflects the model's technical competence to perform as designed under a wide variety of circumstances that may threaten its performance. Ensuring robustness may include validating that the algorithm is reliable (outputs are as expected) and resilient against inputs that may alter the model's stability or produce outputs outside the design's boundaries (Lekadir et al., 2024; NIST, 2023). Ensuring robustness means testing the model's mathematical adherence and seeing if the outputs achieve the intended purposes effectively and safely (Hagendorff, 2019; NIST, 2023; The White House, 2022). The robustness testing would include stress testing, or adversarial testing, to see if the model is resilient to unexpected inputs or bad actor activities (AI HLEG, 2019; Cassidy et al., 2022; Hagendorff, 2019). Since AI does not return the same output each time, traditional testing is not enough; it needs to be tested repeatedly until it is well understood under a wide range of conditions (AI HLEG, 2019; Pichai, 2018). As described in the literature, a model found to be robust under adverse conditions is more ethical than one that is not tested to ensure it is effective, valid, and safe for users.

AI Ethical Framework: Moral Agency

The moral agency (human agency) theme includes related ideas such as avoiding harm, ensuring social benefit, non-maleficence, and respect for human choices. These ideas recognize that AI systems affect not only the immediate user but may also indirectly affect a wide range of other stakeholders (AI HLEG, 2019; Jang et al., 2022). Ethical moral agency focuses on how the user and the wider community are affected by AI choices and outcomes, with or without their knowledge, directly or indirectly, possibly shaping their reality, decisions, and views (Siau & Wang, 2020). Human autonomy, safety, and the well-being of persons apply broadly to the treatment of humans, with health receiving more attention likely because AI use in healthcare is skyrocketing and because of the long tradition of healthcare putting "do no harm" forward in decisions (Arcadi, 2025; Badal et al., 2023; Franco D'Souza et al., 2025; Ho et al., 2020). Google's AI principles take this a bit further, aiming for AI systems to be socially beneficial, not just avoiding harm (Croak, 2023; Pichai, 2018). Those who focus on this aspect of AI ethics emphasize that human beings should remain free to decide for themselves which tools are involved in their life events, including technology. This would include knowing when AI is involved and the ability to opt out of AI use in their affairs (The White House, 2022).

AI Ethical Framework: Value alignment

The value alignment theme represents a cluster of related ideas in the AI ethics literature that encourage AI capabilities to incorporate human values. The value alignment theme would consider questions such as whether AI displaces labor, violates social norms, or disrupts collective well-being. While this theme carries overtones of the moral agency theme above, it asks AI to make the values used in the design and deployment of the AI system explicit. An AI system with explicitly stated collective values would reveal the weighting of priorities and values in tuning the model. Value alignment might be characterized as "society-in-the-loop" in addition to the moral agency of "human-in-the-loop." AI developers would need to actively engage domain experts and diverse stakeholders to fully explore overlapping and consensus values. This challenge is non-trivial, as who spots and decides what are collectively agreed-upon human values has not been articulated and has not historically been a concern for technology designers. We have just started to forge collective AI values, setting up social boundaries with respect to AI. A system with a blind spot for social norms, or perceived as lacking respect for human values, may create unique resistance to adoption and use.

AI Ethical Framework: Technology misuse

Bad actors can use AI to orchestrate a terrifying range of harms. Articulating all the possible harm could fill an entire paper. The dangers include intentional abuse such as deception, AI attacks, social manipulations, impersonations, and misinformation (Bontridder & Pouillet, 2021; Brundage et al., 2018; Habbal et al., 2024; Hirsch et al., 2024; Mayopu et al., 2023; Pichai, 2018). It does not take much imagination to suggest that bad-actor AI can filter for those with cognitive decline, target, and manipulate them. AI can appeal to individuals' unconscious or discern patterns in individuals and treat them differently (Hirsch et al., 2024). If AI systems are involved in deciding to deploy a lethal weapon system, AI could trigger a change in the world order (Habbal et al., 2024). AI is so new that it may have a host of not-yet-considered abuses (Sharma, 2024). In sum, ethical AI engineers must take misuse seriously, thoughtfully consider foreseeable accidental or deliberate AI abuse, and build guardrails (Brundage et al., 2018). Without this, the perceived usefulness and ease of use of AI, key pillars in technology adoption, may be offset by concerns of misuse (Liu & Tao, 2022).

AI Ethical Framework: Trust

AI abilities exceed prior technologies, such as autonomous agency, mimicking human behavior, and training itself, introducing a host of new technology risks that can dampen users' trust (Cabiddu et al., 2022; Glikson & Woolley, 2020). This is a critical issue as adoption of technology by organizations, and AI in particular, depends to a large degree on workers' trust (Lippert & Davis, 2006; Omrani et al., 2022). Unlike prior technologies with automated processing following program rules, AI adjusts its behavior based on inputs (Cabiddu et al., 2022; Kaplan & Haenlein, 2019). An evolving nondeterministic algorithm, a "black box" that recodes itself and offers mysterious outputs, makes it difficult for users to discern the competence of the AI model, making trust more difficult (Bao et al., 2023; Dikmen & Burns, 2022; Kaplan & Haenlein, 2019; Sargent et al., 2024). Trust in AI platforms is not limited to questions about model competence and reliability. Other trust-related concerns include worries about replacing humans, whether the AI model designers had users' best interests in mind, confidence that the AI product is safe and helpful, and how unintended consequences are addressed and by whom (Hengstler et al., 2016; Liu & Tao, 2022; Sharma, 2024). Further, trust in technology is not a stable perception. Recent work suggests that users' trust in AI increases as their use of AI evolves and as their knowledge of the algorithm deepens (Cabiddu et al., 2022; Chi et al., 2021). Some suggest that human-like features of AI can increase trust as perceptions over time as the AI platform evolves into "a loyal pet" (Omrani et al., 2022, p. 1).

AI Ethical Framework: Leadership

Digital leadership and digital transformation drive firm performance, generally through talent management (Koochang et al., 2023; Senadjki et al., 2024). Notably, AI systems are transforming how leaders manage, not just digitizing processes (Hougaard et al., 2024; Kollmann et al., 2023; Madanchian et al., 2024; Zaidi et al., 2025). For instance, AI may augment leadership and meet workers' need for autonomy, mastery, and relatedness, essentially providing the workers with an individualized AI coach, reshaping the world of work (Gibbard et al., 2026; Madanchian et al., 2024; Quaquebeke & Gerpott, 2023). While AI may relieve middle and lower managers from mundane supervisory tasks, research on how AI might impact team dynamics is still limited (Fengkuo et al., 2025; Madanchian et al., 2024). Similar to a smartwatch providing suggestions, reminders, and positive feedback, you can imagine how AI might nudge employees to align their activities with leadership's initiatives (Quaquebeke & Gerpott, 2023; Ren et al., 2024). Regardless of the specific AI practices and policies adopted, leaders need to communicate a well-defined strategic vision for AI, provide workers with a blueprint for change, and train talent to enable the transition to human-AI teams (Bevilacqua et al., 2025; Fengkuo et al., 2025).

AI Frameworks -- Core Principles

In this section, we list and define each framework's core principles based on the literature review.

Data Responsibility

1. **Human Oversight:** AI may be used irresponsibly without human oversight. (Deloitte AI Institute, 2024) (AI HLEG, 2019) (de-Lima-Santos et al., 2024)
2. **Stakeholder Assessment:** AI may be used irresponsibly without input from diverse stakeholders. (AI HLEG, 2019)
3. **Model Assessment:** AI may not perform as intended without rigorous testing and monitoring.
4. **Harm Alerts:** AI potential harms or unintended consequences may go unnoticed because there are no built-in mechanisms to alert users and developers. (Pichai, 2018)
5. **Usage Monitoring:** AI could be misused or adapted for harmful purposes without continuous monitoring. (Pichai, 2018)

Accountability

1. **Commitment to Stakeholders:** AI systems' lack of accountability to all stakeholders, including individuals, communities, and society. (Badal et al., 2023).
2. **Transparency and Notice:** The lack of clear, timely, and accessible notice provided by AI actors to external/public stakeholders regarding decisions made by AI systems. (The White House, 2022)
3. **Beyond Compliance:** AI systems may meet minimum regulatory requirements and industry standards yet still lack a proactive, accountable approach to AI development and deployment. (Hirsch et al., 2024)
4. **Power and Constraints:** The potential for unchecked power conferred upon organizations by AI systems and the absence of clearly established constraints and limitations on their use. (Hirsch et al., 2024)
5. **Independent Evaluation:** The absence of regular assessments by independent third-party organizations evaluating AI systems for their adherence to model safety, ethical principles, and social impact. (The White House, 2022), (Makridis et al., 2024)

Data Privacy

1. **Consent:** AI systems collect and use my personal data without explicit notice and consent. (Pichai, 2018) (de-Lima-Santos et al., 2024) (The White House, 2022)

2. **De-identification:** AI systems are not always designed to de-identify (anonymize/scrub) personal data to protect privacy. (Hirsch et al., 2024)
3. **Re-identification:** The potential risks of re-identification (unmask/uncover), where one's identity could be revealed from seemingly anonymous data. (Hirsch et al., 2024)
4. **Data security/protection:** The security and protection of my personal data when AI systems process it. (IBM, 2024) (NIST, 2023)
5. **Transparency/control:** AI systems are not always transparent about how they use personal data, and people do not always have control over this use. (Pichai, 2018)

Fairness

1. **Unwanted bias:** AI systems perpetuating or amplifying biases, leading to unfair treatment of individuals. (Lekadir et al., 2024)
2. **Vulnerable groups:** AI systems harming vulnerable populations, such as older people, children, or people with disabilities. (Jang et al., 2022).
3. **"Discrimination-aware" data mining:** AI systems making decisions based on sensitive attributes that may be designed to discriminate intentionally. (Jang et al., 2022).
4. **Disparate impact:** AI systems may impact certain groups disparately, even if they are not explicitly targeted. (Hirsch et al., 2024) (The White House, 2022)

Explainability

1. **Explicability:** AI systems are not explainable, meaning their decision-making processes are not transparent to users. (Deloitte AI Institute, 2024) (de-Lima-Santos et al., 2024)
2. **Understandability:** AI systems may not be understandable to users, so they may be unable to comprehend the reasons behind the system's outputs and actions. (Lekadir et al., 2024)
3. **Opacity or black box:** Increasing use of opaque or black-box AI systems, which hide their inner workings from human scrutiny. (Hirsch et al., 2024) (Taylor & Taylor, 2021)
4. **Traceability:** I am concerned that AI systems may not be traceable, preventing users from following the steps that led to a particular decision or output. (Lekadir et al., 2024)
5. **Interpretability:** AI developers will not enable users to understand the capabilities and limitations of the AI tool, making it difficult for them to decide the risk or suitability of use. (Lekadir et al., 2024) (NIST, 2023) (Croak, 2023)

Transparency

1. **Transparency and Accountability:** The lack of transparency in AI decision-making processes, particularly regarding the algorithms used and the rationale behind their outputs. (Jang et al., 2022) (Lekadir et al., 2024) (Makridis et al., 2024)
2. **Traceability and Auditability:** The importance of AI systems having a clear audit trail that allows for the tracking and verification of their actions and decisions. (Lekadir et al., 2024)
3. **Predictability and Control:** I am concerned about how AI systems may not be designed to be predictable and controllable, with clear limitations on their autonomy. (Jang et al., 2022) (Lekadir et al., 2024) (Makridis et al., 2024)
4. **Public Oversight:** The potential lack of public oversight in the development and deployment of AI systems, including the public's ability to challenge and scrutinize the decisions and actions of these systems.

Robustness

1. **Validation:** AI may not undergo rigorous validation processes, leading to inaccurate, ineffective, and unreliable results. (Lekadir et al., 2024) (NIST, 2023)

2. **Model resilience/reliability:** AI models may not be designed and continuously tested to withstand potential failures or attacks, making them vulnerable to disruptions or exploitation. (Lekadir et al., 2024) (NIST, 2023)
3. **Adversarial testing:** AI models may not be subjected to adversarial testing, potentially leaving them vulnerable to malicious actors. (AI HLEG, 2019)
4. **Built and tested for safety:** Safety may not be a primary concern during the development and testing of AI systems. (Hagendorff, 2019) (Pichai, 2018)
5. **Faulty algorithms:** The potential for faulty algorithms to lead to unintended consequences. (Hirsch et al., 2024)

Moral Agency

1. **Avoidance of Harm:** AI systems may cause harm to individuals or society, directly or indirectly. (AI HLEG, 2019)
2. **Social Benefit:** The degree to which AI systems may not be developed or used in ways that are ethical and that benefit society as a whole. (Pichai, 2018)
3. **Respect for Human Autonomy:** AI systems may erode human choice or decision-making abilities. (AI HLEG, 2019)
4. **Human Freedom:** AI systems may limit human freedom and the ability to make independent choices. (AI HLEG, 2019)
5. **Common Good:** AI systems may not be developed and used for the common good of all people. (Pichai, 2018)

Value Alignment

1. **Aligning AI capabilities with human values:** AI systems may not always be designed and developed to prioritize human values. (Deloitte AI Institute, 2024)
2. **Non-maleficence:** AI designs and implementations may not adhere to the principle of non-maleficence (“do no harm”) and could negatively affect individuals or society. (Jang et al., 2022)
3. **Violating internationally accepted norms:** AI systems may be used in ways that violate internationally accepted norms and standards. (Pichai, 2018)
4. **Displacement of labor:** The potential for AI systems to displace human labor and exacerbate social inequalities. (Hirsch et al., 2024)

Technology Misuse

1. **Deception or Masking of True Authors or Sources:** AI systems that deceive people by hiding the true source of information. (Brundage et al., 2018)
2. **AI Attackers and AI Defenders:** The potential for AI systems to be used by malicious actors to attack individuals or organizations, and the difficulty of developing effective AI defenses. (Brundage et al., 2018)
3. **Difficulties in Determining Truthful Public Debates:** The increasing difficulty of distinguishing between truthful and false information in public debates, especially when AI-generated content is involved. Brundage et al., 2018)
4. **Misinformation:** The spread of misinformation and disinformation through AI-generated content. (Bontridder & Poulet, 2021)
5. **Intentionally Harmful AI:** Developing AI systems specifically designed to cause harm, such as autonomous weapons or cyberattack tools. (Hirsch et al., 2024)

Trust

1. **Competence:** AI systems generally perform tasks effectively and reliably. (NIST, 2023)

2. **Integrity:** I believe AI systems are generally honest and do not intentionally mislead me. (Makridis et al., 2024)
3. **Benevolence:** AI systems are generally developed and used with the best interests of humans in mind. (AI HLEG, 2019)
4. **Confidence:** AI system designers intend to provide a safe, helpful product and will adapt and address unintended consequences transparently. (AI HLEG, 2019)

Leadership

1. **Develop a strategic vision for AI:** Leadership is involved in identifying opportunities to leverage AI for competitive advantage, setting clear goals, and aligning AI initiatives with the overall business strategy. (Adapted for AI from Koohang et al., 2023)
2. **Communicate transparently:** Leadership communicates openly and honestly with employees about the organization's AI strategy, addressing concerns and fostering trust. (Adapted for AI from Koohang et al., 2023)
3. **Invest in AI talent and skills:** Leadership provides training and development opportunities for existing employees to adapt to AI-driven roles. (Adapted for AI from Koohang et al., 2023)
4. **Prioritize ethical considerations:** AI poses significant ethical challenges, including bias, privacy concerns, and the potential for job displacement. Leadership prioritizes responsible AI practices and proactively addresses these issues to ensure ethical development and deployment. (Adapted for AI from Koohang et al., 2023)

Methodology

Instrument

The instrument for this study was designed based on the AI ethics frameworks discussed above, grouped into 12 constructs/frameworks, each with its associated items/core principles. Users were asked to express their concerns about the core principles within each construct. The constructs are data responsibility, accountability, data privacy, fairness, explainability, transparency, robustness, moral agency, value alignment, technology misuse, leadership, and trust. The instrument used a Likert-type scale with the following scoring strategy: Strongly Agree = 5, Agree = 4, Neutral = 3, Disagree = 2, and Strongly Disagree = 1.

Procedure and Sample

The study was granted exempt status by the originating institution's Institutional Review Board (IRB). Data were collected via a third-party internet survey firm, yielding 244 usable responses from diverse organizations across the United States, spanning nine geographic regions. The East North Central division has the most participants (44). Significant participation also appears in the South Atlantic (38), Pacific (36), and Middle Atlantic (33) regions. Mid-range representation comes from the West South Central (25), West North Central (21), and East South Central (19) areas. Rounding out the demographics are the Mountain and New England regions, contributing 16 and 9 participants, respectively, ensuring a comprehensive nationwide perspective. The gender distribution comprises 152 females and 92 males. Age demographics are dominated by the 30–44 group (n=83), followed by 45–60 (n=64), 18–29 (n=53), and 60 and above (n=44). Professionally, the group is diverse, with the largest segment at the Intermediate job level (n=78), followed by Entry Level (n=56), Middle Management (n=55), and high-level leadership, including Owners and Executives (n=32) and Senior Management (n=23). Regarding AI integration, organizational priority is largely split between Moderate (n=106) and Low (n=99), with only 39 participants reporting a High priority. This aligns with reported skill levels, where a majority are Limited (n=92) or Somewhat (n=82) proficient, while only 22 are Highly proficient. Daily AI usage is distributed almost evenly across the

spectrum, though the most frequent response indicates users are Very likely ($n=55$) to engage with the technology, closely followed by those who are Slightly likely ($n=53$) or Not at all likely ($n=52$).

Data Analysis

Data were analyzed using Exploratory Factor Analysis (EFA) in SPSS™ (Version 28). EFA is used to identify the underlying constructs for a set of measured variables. Following the framework established by Mertler and Vannatta (2001), the analysis included four preliminary tests to assess data suitability before proceeding to the final Principal Component Analysis (PCA) with Varimax rotation. To satisfy the criteria for factor retention, several statistical thresholds must be met: the Kaiser-Meyer-Olkin (KMO) index for sampling adequacy must exceed 0.60, and Bartlett's Test of Sphericity must be statistically significant ($p < .05$). Under the Kaiser Criterion, factors with eigenvalues greater than 1.0 are retained—a method considered most reliable when the sample size exceeds 200, variables range between 10 and 40, or communalities are approximately .60 or higher. Furthermore, the retained components should collectively account for an acceptable portion of the total variance, ideally close to 70%, and a Scree Plot should be used to visually identify components above the "breakpoint" before the eigenvalues level off (Mertler & Vannatta, 2001). After confirming the data's suitability through preliminary tests, a Principal Component Analysis (PCA) with Varimax rotation was conducted to answer the research question. This method identifies distinct components and enhances factor loadings for better interpretability. To verify the internal consistency of items within each identified component, Cronbach's alpha was calculated; following Mertler and Vannatta (2001), a reliability coefficient of or higher was used as the acceptance threshold.

Results

Initial Exploratory Factor Analysis (EFA) revealed cross-loadings across several items. These items were from the following constructs: explainability, transparency, robustness, moral agency, value alignment, and technology misuse. Following the removal of these items, the data were reanalyzed to confirm the final factor solution. The results are presented below. Sampling adequacy was assessed using the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity. For the data to be considered suitable, the KMO index must exceed 0.6, and Bartlett's test must be significant ($< .05$). These tests ensure that the variables are sufficiently intercorrelated and that the sample size is adequate for the intended multivariate analysis. The Kaiser-Meyer-Olkin value was 0.939, indicating that the sampling adequacy criterion was met. Bartlett's test of sphericity ($\text{Chi-Squared} = 8121.416$, $df = 1035$, $p < .001$) indicated that the data collected were suitable for analysis. Furthermore, all communalities values were approximately .60 or greater.

The variance test (retained for all components, including the initial eigenvalues and the rotation sums of squared loadings) showed that the total retained variance accounted for 68% across eight (8) components, which is within the acceptable variability threshold. Furthermore, the Scree Plot Analysis visually identified components situated above the "breakpoint" before the eigenvalues level off. Table 1 presents the results of Principal Component Analysis (PCA) with Varimax rotation, conducted to address the research question. A cut-off point of .45 was used, which is generally considered acceptable and practically significant and is often used to denote moderate to strong factor loadings (Jolliffe & Cadima, 2016). As shown, 8 components were retained from the original 12. Also, see Appendix 1.

Component 1 retained all five items from the Data responsibility construct and all five from the Accountability construct, resulting in a single component. The items for this component are as follows. Data responsibility_1: Human Oversight; Data responsibility_2: Stakeholder Assessment; Data responsibility_3: Model Assessment; Data responsibility_4: Harm Alerts; Data responsibility_5: Usage Monitoring; Accountability_1: Commitment to Stakeholders; Accountability_2: Transparency and Notice;

Accountability_3: Beyond Compliance; Accountability_4: Power and Constraints; and Accountability_5: Independent Evaluation. The reliability test for this component was .919.

Component 2 retained 3 items from the Robustness construct and 4 items from the Technology misuse construct, forming a single component. The items for this component are as follows. Robustness_1: Validation; Robustness_2: Model resilience/reliability; Robustness_3: Adversarial testing; Technology misuse_1: Deception or Masking of True Authors/ Sources; Technology misuse_2: AI Attackers and AI Defenders; Technology misuse_3: Difficulties in Determining Truthful Public Debates; and Technology misuse_4: Misinformation. The reliability test for this component was .910.

Component 3 retained all five items from the Data privacy construct. Data privacy_1: Consent; Data privacy_2: De-identification; Data privacy_3: Re-identification; Data privacy_4: Data security/protection; and Data privacy_5: Transparency/control. The reliability test for this component was .874.

Component 4 retained 3 of the 4 items from the Explainability construct and 3 of the 4 items from the Transparency construct, resulting in a single component. The items for this component are as follows. Explainability_2: Understandability; Explainability_4: Traceability; Explainability_5: Interpretability; Transparency_1: Transparency and Accountability; Transparency_2: Traceability and Auditability; and Transparency_3: Predictability and Control. The reliability test for this component was .889.

Component 5 retained all 4 items from the Fairness construct. The items are as follows. Fairness_1: Unwanted bias; Fairness_2: Vulnerable groups; Fairness_3: “Discrimination-aware” data mining; and Fairness_4: Disparate impact. The reliability test for this component was .893.

Component 6 retained 4 of the 5 items from the Moral Agency construct and 2 of the 4 items from the Value Judgment construct. The items are as follows. Moral agency_1: Avoidance of Harm; Moral agency_3: Respect for Human Autonomy; Moral agency_4: Human Freedom; Moral agency_5: Common Good; Value alignment_2: Non-maleficence; and Value alignment_4: Displacement of labor. The reliability test for this component was .875

Component 7 retained all four items from the Leadership construct. The items are as follows. Leadership_1: Develop a strategic vision for AI; Leadership_2: Communicate transparently; Leadership_3: Invest in AI talent and skills; and Leadership_4: Prioritize ethical considerations. The reliability test for this component was .810.

Component 8 retained all four items from the Trust construct. The items are as follows. Trust_1: Competence; Trust_2: Integrity; Trust_3: Benevolence; and Trust_4: Confidence. The reliability test for this component was .843.

Table 1. Rotated Component Matrix

		Component							
		1	2	3	4	5	6	7	8
1	Data responsibility 1	0.594	0.236	0.237	0.199	-0.041	0.310	-0.026	-0.160
	Data responsibility 2	0.738	0.021	0.030	0.199	0.009	0.141	0.112	-0.026
	Data responsibility 3	0.683	0.184	0.195	0.108	0.207	0.071	-0.017	0.040
	Data responsibility 4	0.711	0.241	0.210	-0.016	0.224	0.194	-0.023	-0.022
	Data responsibility 5	0.586	0.241	0.389	0.040	0.148	0.287	0.050	-0.066
	Accountability 1	0.637	0.150	0.383	0.169	0.289	0.052	0.028	-0.019
	Accountability 2	0.642	0.141	0.216	0.336	0.208	0.090	0.014	0.062
	Accountability 3	0.617	0.261	0.120	0.232	0.264	0.106	-0.052	0.061

	Accountability 4	0.479	0.379	0.238	0.336	0.223	0.152	-0.105	0.006
	Accountability 5	0.488	0.156	0.356	0.274	0.328	0.169	-0.030	-0.031
2	Robustness 1	0.212	0.562	0.214	0.284	0.344	0.221	-0.023	-0.103
	Robustness 2	0.253	0.516	0.125	0.092	0.329	0.373	-0.038	-0.063
	Robustness 3	0.173	0.650	0.075	0.231	0.295	0.313	0.018	0.011
	Technology misuse 1	0.331	0.457	0.285	0.400	0.252	0.087	0.053	-0.040
	Technology misuse 2	0.225	0.646	0.122	0.404	0.096	0.250	0.059	-0.114
	Technology misuse 3	0.368	0.602	0.250	0.135	0.295	0.123	0.027	-0.006
	Technology misuse 4	0.311	0.701	0.288	0.076	0.220	0.158	-0.015	-0.095
3	Data privacy 1	0.254	0.029	0.765	0.131	0.146	0.222	-0.005	-0.110
	Data privacy 2	0.215	0.113	0.546	0.341	0.123	0.222	0.065	-0.064
	Data privacy 3	0.263	0.215	0.576	0.394	0.200	0.037	-0.011	-0.064
	Data privacy 4	0.293	0.237	0.705	0.230	0.146	0.131	-0.100	-0.059
	Data privacy 5	0.369	0.279	0.666	0.136	0.101	0.049	-0.119	-0.002
4	Explainability 2	0.163	0.192	0.176	0.708	0.169	0.204	-0.049	0.030
	Explainability 4	0.385	0.199	0.129	0.496	0.408	0.120	0.018	-0.042
	Explainability 5	0.239	0.337	0.209	0.627	0.107	0.182	0.039	-0.041
	Transparency 1	0.306	0.050	0.211	0.681	0.264	0.226	0.060	-0.074
	Transparency 2	0.181	0.149	0.386	0.548	0.351	0.151	0.059	-0.004
	Transparency 3	0.153	0.324	0.376	0.458	0.364	0.182	-0.011	0.054
5	Fairness 1	0.285	0.227	0.250	0.253	0.698	0.087	0.006	0.022
	Fairness 2	0.190	0.230	0.232	0.265	0.595	0.266	-0.028	-0.039
	Fairness 3	0.242	0.335	0.181	0.196	0.741	0.086	0.112	-0.023
	Fairness 4	0.289	0.229	0.070	0.211	0.718	0.267	0.070	-0.057
6	Moral agency 1	0.377	0.238	0.179	0.276	0.180	0.477	0.072	-0.161
	Moral agency 3	0.238	0.294	0.064	0.203	0.167	0.666	-0.006	0.039
	Moral agency 4	0.203	0.130	0.185	0.233	0.113	0.809	0.012	0.039
	Moral agency 5	0.324	0.421	0.173	0.308	0.205	0.465	-0.072	-0.058
	Value alignment 2	0.416	0.290	0.313	0.225	0.161	0.479	0.043	-0.028
	Value alignment 4	0.056	0.248	0.414	0.036	0.308	0.495	0.001	0.048
7	Leadership 1	0.029	-0.032	-0.055	-0.068	0.089	0.151	0.725	0.325
	Leadership 2	-0.017	0.019	0.081	-0.075	0.010	-0.054	0.846	0.142
	Leadership 3	-0.102	-0.046	0.003	-0.023	0.035	-0.050	0.819	0.201
	Leadership 4	0.174	0.127	-0.099	0.245	-0.034	-0.003	0.665	0.072
8	Trust 1	-0.017	0.228	0.087	-0.075	-0.047	-0.102	0.080	0.795
	Trust 2	0.014	-0.106	-0.074	-0.066	-0.009	0.014	0.200	0.846
	Trust 3	-0.006	-0.194	-0.131	0.131	-0.082	0.071	0.371	0.727
	Trust 4	-0.058	-0.221	-0.128	-0.008	0.070	0.026	0.430	0.682

Discussion

The results of this Exploratory Factor Analysis (EFA) suggest that, although the theoretical framework for AI ethics is often discussed through 12 distinct lenses, users perceive these concerns through an 8-factor, more consolidated structure. The high KMO value (.939) and significant Bartlett's test confirm that the instrument is robust, and the sample size was more than adequate for this multivariate analysis. Notably, the collapse of several original constructs into shared components indicates substantial conceptual overlap among stakeholders. For instance, Data Responsibility and Accountability merged into a single powerhouse component (Component 1) with a reliability of .919. This suggests that users do not differentiate between responsibility for data and the systemic structures that hold an organization accountable; rather, they view

them as a single "stewardship" dimension. Similarly, the blending of Explainability and Transparency (Component 4) confirms that, in a practical organizational context, the ability to understand a model (interpretability) is seen as inseparable from the visibility of its processes (traceability). The consolidation of Robustness and Technology Misuse (Component 2) is also telling; it implies that users view an AI system's technical resilience primarily through the lens of its susceptibility to malicious exploitation or misinformation. These findings highlight that, while academics may draw fine lines between these twelve ethical pillars, practitioners and users perceive a more integrated landscape of risk and responsibility.

Implications

The primary implication of this study is the development of a validated, streamlined instrument for organizations to assess ethical concerns related to AI. By reducing the complexity from 12 constructs to 8 reliable components, the survey becomes more efficient for deployment without sacrificing depth. For organizational leaders, the emergence of Leadership (Component 7) and Trust (Component 8) as standalone factors is significant. It underscores that ethical AI is not merely a technical challenge of "fixing the algorithm," but a cultural one that requires a clear strategic vision and the cultivation of benevolence and integrity. For example, this finding is paramount to C-suite leadership in the healthcare sector, who are consistently considering the ethical implications of various aspects of AI involvement with healthcare organizational decision-making and patient care (Giovanola & Tiribelli, 2023).

Furthermore, the merging of Moral Agency and Value Alignment (Component 6) suggests that users view the "common good" and "non-maleficence" as a unified ethical compass. Organizations should use these 8 components as a roadmap for governance. Instead of siloed teams working on "privacy" vs. "fairness," these results suggest that cross-functional teams should address these dimensions collectively, through a "Transparency and Explainability" task force, to better align with how stakeholders naturally perceive AI risks.

Future Research

While this study provides a strong empirical foundation, the removal of several items due to cross-loadings (including "Public Oversight" and "Intentionally Harmful AI") suggests that certain nuances of AI ethics remain difficult to isolate. Future research should employ Confirmatory Factor Analysis (CFA) on a new sample to validate this 8-factor model and assess its stability across industries, such as healthcare and finance, where definitions of "fairness" or "robustness" may differ. Additionally, longitudinal studies could examine how these 8 components evolve as AI literacy in the general public increases. For example, will users eventually learn to distinguish between "Transparency" and "Explainability" as they become more familiar with "black box" technologies? Future research should be mindful that public views on AI ethics are expected to evolve over time and become more nuanced as general AI literacy increases. As individuals gain familiarity with complex "black box" technologies, they will likely develop a sharper understanding of these issues and eventually learn to distinguish between closely related concepts such as "Transparency" and "Explainability." Furthermore, the practical understanding of core ethical principles, such as "fairness" and "robustness," may shift as they are applied and adapted to the distinct contexts of different industries over time. Finally, future studies should assess the predictive validity of this instrument, specifically whether high scores on these 8 ethical components correlate with higher user adoption rates or long-term brand loyalty among organizations deploying AI solutions.

Conclusion

Empirical validation of this 12-construct framework reveals a critical shift in how users perceive AI ethics: they do not see 12 isolated silos but rather 8 integrated dimensions of institutional responsibility. The high internal consistency across all components (Cronbach's alpha values ranging from .810 to .919) indicates that this instrument is a highly reliable tool for organizations seeking to audit their ethical standing. The seamless merging of Data Responsibility and Accountability suggests that, for the end user, the "how" (data management) and the "who" (organizational oversight) are two sides of the same coin. Ultimately, this study bridges the gap between theoretical ethical guidelines and practical user concerns. By identifying Trust and Leadership as independent pillars, the research shows that technical robustness alone is insufficient for AI adoption. An organization can have a "safe" algorithm, but without a strategic vision and demonstrated benevolence, user concerns will persist. This 8-factor model offers a streamlined, validated roadmap for developers and policymakers to build AI systems that are not only technically sound but also socially and morally aligned with human expectations.

References

- AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. High-level expert group on artificial intelligence.
- Arcadi, P. (2025). Nursing leadership and artificial intelligence ethics: Safeguarding relationships and values. *Nursing Ethics*, 32(8), 2468–2476. <https://doi.org/10.1177/09697330251366599>
- Badal, K., Lee, C. M., & Esserman, L. J. (2023). Guiding principles for the responsible development of artificial intelligence tools for healthcare. *Communications Medicine*, 3(1), 47–47. <https://doi.org/10.1038/s43856-023-00279-9>
- Bao, Y., Gong, W., & Yang, K. (2023). A Literature Review of Human–AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems (Basel)*, 11(9), 442. <https://doi.org/10.3390/systems11090442>
- Becker, S. J., Nemat, A. T., Lucas, S., Heinitz, R. M., Klevesath, M., & Charton, J. E. (2023). A Code of Digital Ethics: Laying the foundation for digital ethics in a science and technology company. *AI & Society*, 38(6), 2629–2639. <https://doi.org/10.1007/s00146-021-01376-w>
- Bergquist, M., Rolandsson, B., Gryska, E., Laesser, M., Hoefling, N., Heckemann, R., Schneiderman, J. F., & Björkman-Burtscher, I. M. (2024). Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. *European Radiology*, 34(1), 338–347. <https://doi.org/10.1007/s00330-023-09967-5>
- Bevilacqua, S., Masárová, J., Perotti, F. A., & Ferraris, A. (2025). Enhancing top managers' leadership with artificial intelligence: Insights from a systematic literature review. *Review of Managerial Science*, 19(9), 2899–2935. <https://doi.org/10.1007/s11846-025-00836-7>
- Bontridder, N., & Pouillet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3. ProQuest Central; ProQuest One Academic (2601979788). <https://doi.org/10.1017/dap.2021.20>
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó.,

- Beard, S., Belfield, H., Farquhar, S., ... Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. <https://doi.org/10.48550/arxiv.1802.07228>
- Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, 40(5), 685–706. <https://doi.org/10.1016/j.emj.2022.06.001>
- Cassidy, B., Hittner, R., Crowley, B., & Bowman, Z. (2022). *An auditor's mindset in an AI-driven world*. Deloitte. <https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/advancing-human-ai-collaboration.html>
- Chi, O. H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, 118, 106700. <https://doi.org/10.1016/j.chb.2021.106700>
- Christoforaki, M., & Beyan, O. (2022). AI Ethics-A Bird's Eye View. *Applied Sciences*, 12(9), 4130. <https://doi.org/10.3390/app12094130>
- Colin Clemente Jones. (2023). Systematizing discrimination: AI vendors and Title VII enforcement. *University of Pennsylvania Law Review*, 171(1), 235–266.
- Croak, M. (2023). *Google Research, 2022 & beyond: Responsible AI*. https://research.google/blog/google-research-2022-beyond-responsible-ai/?_gl=1*d8xwck*_ga*MTg3MDI4MjE1NC4xNzMxNzcyMDc2*_ga_KFG60X%E2%80%A6
- de-Lima-Santos, M.-F., Yeung, W. N., & Dodds, T. (2024). Guiding the way: A comprehensive examination of AI guidelines in global media. *AI & Society*. <https://doi.org/10.1007/s00146-024-01973-5>
- Deloitte AI Institute. (2024). *Toward humanity's brightest future with generative AI*. www.deloitte.com/us/AIInstitute
- Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20(1), 1–3. ProQuest Central; ProQuest One Academic (2002037084). <https://doi.org/10.1007/s10676-018-9450-z>
- Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162, 102792. <https://doi.org/10.1016/j.ijhcs.2022.102792>
- Fengkuo, S., Yijia, Z., Comite, U., Badulescu, A., & Badulescu, D. (2025). Artificial Intelligence-Supported Leadership: A Catalyst for Team Excellence in China's Fast-Moving Consumer Goods Industry. *Journal of Organizational and End User Computing*, 37(1), 1–29. <https://doi.org/10.4018/JOEUC.379769>
- Franco D'Souza, R., Surapaneni, K. M., P, S., Regupathy, A., Mathew, M., Mishra, V., Kalaimathi, A. G., Sekkizhar, G., Tandon, R., Louis Palatty, P., & Mady, V. (2025). Convergence of Diverse Expertise: A Multidisciplinary Training on the Ethics of Artificial Intelligence in Healthcare Technology and Research. *Journal of Academic Ethics*, 23(3), 885–899. <https://doi.org/10.1007/s10805-024-09593-w>

- Gibbard, K., Gill, H., Powell, D., & Hausdorf, P. A. (2026). Explain it to me like I'm five: Harnessing the power of explanations to increase trust in workplace generative AI. *Behaviour & Information Technology*, 45(1), 40–58. <https://doi.org/10.1080/0144929X.2025.2506664>
- Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society*, 38(2), 549–563. <https://doi.org/10.1007/s00146-022-01455-6>
- Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *The Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Gong, D. (2026). Closing the responsibility gap: Allocating responsibility according to prerequisite control and expectations for personal benefits. *Ethics and Information Technology*, 28(1), 8. <https://doi.org/10.1007/s10676-025-09883-w>
- Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications*, 240, 122442. <https://doi.org/10.1016/j.eswa.2023.122442>
- Hagendorff, T. (2019). The Ethics of AI Ethics—An Evaluation of Guidelines. *arXiv.Org*. ProQuest Central; ProQuest One Academic (2189902325). <https://doi.org/10.1007/s11023-020-09517-8>
- Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting & Social Change*, 105, 105–120. <https://doi.org/10.1016/j.techfore.2015.12.014>
- Higgins, O., Short, B. L., Chalup, S. K., & Wilson, R. L. (2023). Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review. *International Journal of Mental Health Nursing*, 32(4), 966–978. <https://doi.org/10.1111/inm.13114>
- Hirsch, Dennis., Bartley, Timothy., Chandrasekaran, Aravind., Norris, Davon., Parthasarathy, Srinivasan., & Turner, P. Norris. (2024). *Business Data Ethics: Emerging Models for Governing AI and Advanced Analytics* (1st ed. 2024.). Springer International Publishing. <https://doi.org/10.1007/978-3-031-21491-2>
- Ho, C. W. L., Ali, J., & Caals, K. (2020). Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bulletin of the World Health Organization*, 98(4), 263–269. <https://doi.org/10.2471/BLT.19.234732>
- Hougaard, R., Carter, J., & Stembridge, R. (2024). The best leaders can't be replaced by AI. *Harvard Business Review*. <https://hbr.org/2024/01/the-best-leaders-cant-be-replaced-by-ai>
- IBM. (2024). *What is AI ethics?* IBM. www.ibm.com/topics/ai-ethics
- Jang, Y., Choi, S., & Kim, H. (2022). Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (AT-EAI) and analysis of its difference by gender and experience of AI education. *Education and Information Technologies*, 27(8), 11635–11667. <https://doi.org/10.1007/s10639-022-11086-5>

- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Kaplan, A., & 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*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- Köchling, A., & Wehner, M. C. (2023). Better explaining the benefits why AI? Analyzing the impact of explaining the benefits of AI-supported selection on applicant responses. *International Journal of Selection and Assessment*, 31(1), 45–62. <https://doi.org/10.1111/ijsa.12412>
- Kollmann, T., Kollmann, K., & Kollmann, N. (2023). Artificial intelligence: Digital transformation as a leadership task between the chief digital officer and artificial intelligence. *International Journal of Business Science & Applied Management*, 18(1), 79–95. <https://doi.org/10.69864/ijbsam.18-1.172>
- Koohang, A., Sargent, C. S., Zhang, J. Z., & Marotta, A. (2023). Big data analytics: From leadership to firm performance. *Industrial Management + Data Systems*, 123(12), 2976–2996. <https://doi.org/10.1108/IMDS-06-2023-0415>
- Lekadir, K., Feragen, A., Fofanah, A. J., Frangi, A. F., Buyx, A., Lara, A., Porras, A. R., An-Wen, C., Navarro, A., Botwe, B. O., Khanal, B., Beger, B., Wu, C. C., Cintas, C., Langlotz, C. P., Rueckert, D., Mzurikwao, D., Fotiadis, D. I., Zhussupov, D., ... Starmans, M. P. A. (2024). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *arXiv.Org*.
- Liang, L., & Acuna, D. E. (2020). *Artificial mental phenomena: Psychophysics as a framework to detect perception biases in AI models*. 403–412. <https://doi.org/10.1145/3351095.3375623>
- Lippert, S. K., & Davis, M. (2006). A conceptual model integrating trust into planned change activities to enhance technology adoption behavior. *Journal of Information Science*, 32(5), 434–448. <https://doi.org/10.1177/0165551506066042>
- Liu, K., & Tao, D. (2022). The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Computers in Human Behavior*, 127, 107026. <https://doi.org/10.1016/j.chb.2021.107026>
- Madanchian, M., Taherdoost, H., Vincenti, M., & Mohamed, N. (2024). Transforming Leadership Practices through Artificial Intelligence. *Procedia Computer Science*, 235, 2101–2111. <https://doi.org/10.1016/j.procs.2024.04.199>
- Makridis, C. A., Mueller, J., Tiffany, T., Borkowski, A. A., Zachary, J., & Alterovitz, G. (2024). From theory to practice: Harmonizing taxonomies of trustworthy AI. *Health Policy OPEN*, 7, 100128. <https://doi.org/10.1016/j.hpopen.2024.100128>
- Mayopu, R. G., Wang, Y.-Y., & Chen, L.-S. (2023). Analyzing Online Fake News Using Latent Semantic Analysis: Case of USA Election Campaign. *Big Data and Cognitive Computing*, 7(2), 81. <https://doi.org/10.3390/bdcc7020081>

- Mertler, C., & Vannatta, R. (2001). *Advanced and multivariate statistical methods: Practical application and interpretation*. Pyrczak Publishing.
- Mökander, J., Sheth, M., Gersbro-Sundler, M., Blomgren, P., & Floridi, L. (2022). Challenges and best practices in corporate AI governance: Lessons from the biopharmaceutical industry. *Frontiers in Computer Science (Lausanne)*, 4. <https://doi.org/10.3389/fcomp.2022.1068361>
- NIST. (2023). *Artificial intelligence risk management framework (AI RM 1.0)*. U.S. Department of Commerce.
- Omrani, N., Riviuccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting & Social Change*, 181, 121763. <https://doi.org/10.1016/j.techfore.2022.121763>
- Pichai, S. (2018). *AI at Google: Our principles*. <https://blog.google/technology/ai/ai-principles/>
- Quaquebeke, N. V., & Gerpott, F. H. (2023). The Now, New, and Next of Digital Leadership: How Artificial Intelligence (AI) Will Take Over and Change Leadership as We Know It. *Journal of Leadership & Organizational Studies*, 30(3), 265–275. <https://doi.org/10.1177/15480518231181731>
- Raman, R., Kowalski, R., Achuthan, K., Iyer, A., & Nedungadi, P. (2025). Navigating artificial general intelligence development: Societal, technological, ethical, and brain-inspired pathways. *Scientific Reports*, 15(1), 8443–22. <https://doi.org/10.1038/s41598-025-92190-7>
- Ren, S., Mawritz, M. B., Greenbaum, R. L., Babalola, M. T., & Wang, Z. (2024). Does Competitive Action Intensity Influence Team Performance via Leader Bottom-Line Mentality? A Social Information Processing Perspective. *Journal of Applied Psychology*, 109(6), 811–828. <https://doi.org/10.1037/apl0001166>
- Saba, C. S., & Pretorius, M. (2024). The impact of artificial intelligence (AI) investment on human well-being in G-7 countries: Does the moderating role of governance matter? *Sustainable Futures*, 7, 100156. <https://doi.org/10.1016/j.sftr.2024.100156>
- Sargent, C., Koohang, A., Floyd, K., & Kilburn, R. (2024). Artificial intelligence: Ethical concerns, trust and risk. *Issues in Information Systems*, 25(2), 71–83. https://doi.org/10.48009/2_iis_2024_106
- Senadjki, A., Au Yong, H. N., Ganapathy, T., & Ogbeibu, S. (2024). Unlocking the potential: The impact of digital leadership on firms' performance through digital transformation. *Journal of Business and Socio-Economic Development*, 4(2), 161–177. <https://doi.org/10.1108/JBSED-06-2023-0050>
- Sharma, S. (2024). Benefits or concerns of AI: A multistakeholder responsibility. *Futures : The Journal of Policy, Planning and Futures Studies*, 157, 103328. <https://doi.org/10.1016/j.futures.2024.103328>
- Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, 31(2), 74–87. <https://doi.org/10.4018/JDM.2020040105>
- Stix, C. (2021). Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Science and Engineering Ethics*, 27(1), 15–15. <https://doi.org/10.1007/s11948-020-00277-3>

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2), 454–475. <https://doi.org/10.3758/s13423-020-01825-5>

The White House. (2022). *Blueprint for an AI bill of rights: Making automated systems work for the American people*.

Wecker, L., Nielsen, J. A., Nasrollahi, K., & Ploug, T. (2026). A review of ethical AI frameworks in product development: Taking stock and moving forward. *Ai and Ethics (Online)*, 6(1). <https://doi.org/10.1007/s43681-025-00952-w>

Zaidi, S. Y. A., Aslam, M. F., Mahmood, F., Ahmad, B., & Raza, S. B. (2025). How Will Artificial Intelligence (AI) Evolve Organizational Leadership? Understanding the Perspectives of Technopreneurs. *Global Business and Organizational Excellence*, 44(3), 66–83. <https://doi.org/10.1002/joe.22275>

Appendix A The Final Instrument with eight (8) components

Component 1: Data Responsibility/Accountability
1. Human Oversight: AI may be used irresponsibly without human oversight.
2. Stakeholder Assessment: AI may be used irresponsibly without input from diverse stakeholders.
3. Model Assessment: AI may not perform as intended without rigorous testing and monitoring.
4. Harm Alerts: AI potential harms or unintended consequences may go unnoticed because there are no built-in mechanisms to alert users and developers.
5. Usage Monitoring: AI could be misused or adapted for harmful purposes without continuous monitoring.
6. Commitment to Stakeholders: AI systems' lack of accountability to all stakeholders, including individuals, communities, and society.
7. Transparency and Notice: The lack of clear, timely, and accessible notice provided by AI actors to external/public stakeholders regarding decisions made by AI systems.
8. Beyond Compliance: AI systems may meet minimum regulatory requirements and industry standards yet still lack a proactive, accountable approach to AI development and deployment.
9. Power and Constraints: The potential for unchecked power conferred upon organizations by AI systems and the absence of clearly established constraints and limitations on their use.
10. Independent Evaluation: The absence of regular assessments by independent third-party organizations evaluating AI systems for their adherence to model safety, ethical principles, and social impact.
Component 2: Data Privacy
1. Consent: AI systems collect and use my personal data without explicit notice and consent.
2. De-identification: AI systems are not always designed to de-identify (anonymize/scrub) personal data to protect privacy.
3. Re-identification: The potential risks of re-identification (unmask/uncover), where one's identity could be revealed from seemingly anonymous data.
4. Data security/protection: The security and protection of my personal data when AI systems process it.
5. Transparency/control: AI systems are not always transparent about how they use personal data, and people do not always have control over this use.
Component 3: Fairness
1. Unwanted bias: AI systems perpetuating or amplifying biases, leading to unfair treatment of individuals.
2. Vulnerable groups: AI systems harming vulnerable populations, such as older people, children, or people with disabilities.
3. "Discrimination-aware" data mining: AI systems making decisions based on sensitive attributes that may be designed to discriminate intentionally.
4. Disparate impact: AI systems may impact certain groups disparately, even if they are not explicitly targeted.
Component 4: Explainability/Transparency
1. Understandability: AI systems may not be understandable to users, so they may be unable to comprehend the reasons behind the system's outputs and actions.
2. Traceability: I am concerned that AI systems may not be traceable, preventing users from following the steps that led to a particular decision or output.
3. Interpretability: AI developers will not enable users to understand the capabilities and limitations of the AI tool, making it difficult for them to decide the risk or suitability of use.
4. Transparency and Accountability: The lack of transparency in AI decision-making processes, particularly regarding the algorithms used and the rationale behind their outputs.
5. Traceability and Auditability: The importance of AI systems having a clear audit trail that allows for the tracking and verification of their actions and decisions.
6. Predictability and Control: I am concerned about how AI systems may not be designed to be predictable and controllable, with clear limitations on their autonomy.
Component 5: Robustness/Technology Misuse
1. Validation: AI may not undergo rigorous validation processes, leading to inaccurate, ineffective, and unreliable results.

2. Model resilience/reliability: AI models may not be designed and continuously tested to withstand potential failures or attacks, making them vulnerable to disruptions or exploitation.
3. Adversarial testing: AI models may not be subjected to adversarial testing, potentially leaving them vulnerable to malicious actors.
4. Deception or Masking of True Authors or Sources: AI systems that deceive people by hiding the true source of information.
5. AI Attackers and AI Defenders: The potential for AI systems to be used by malicious actors to attack individuals or organizations, and the difficulty of developing effective AI defenses.
6. Difficulties in Determining Truthful Public Debates: The increasing difficulty of distinguishing between truthful and false information in public debates, especially when AI-generated content is involved.
7. Misinformation: The spread of misinformation and disinformation through AI-generated content.
Component 6: Moral Agency/Value Alignment
1. Avoidance of Harm: AI systems may cause harm to individuals or society, directly or indirectly.
2. Respect for Human Autonomy: AI systems may erode human choice or decision-making abilities.
3. Human Freedom: AI systems may limit human freedom and the ability to make independent choices.
4. Common Good: AI systems may not be developed and used for the common good of all people.
5. Non-maleficence: AI designs and implementations may not adhere to the principle of non-maleficence (do no harm”) and could negatively affect individuals or society.
6. Displacement of labor: The potential for AI systems to displace human labor and exacerbate social inequalities.
Component 7: Trust
1. Competence: AI systems generally perform tasks effectively and reliably.
2. Integrity: I believe AI systems are generally honest and do not intentionally mislead me.
3. Benevolence: AI systems are generally developed and used with the best interests of humans in mind.
4. Confidence: AI system designers intend to provide a safe, helpful product and will adapt and address unintended consequences transparently.
Component 8: Leadership
1. Develop a strategic vision for AI: Leadership is involved in identifying opportunities to leverage AI for competitive advantage, setting clear goals, and aligning AI initiatives with the overall business strategy.
2. Communicate transparently: Leadership communicates openly and honestly with employees about the organization's AI strategy, addressing concerns and fostering trust.
3. Invest in AI talent and skills: Leadership provides training and development opportunities for existing employees to adapt to AI-driven roles.
4. Prioritize ethical considerations: AI poses significant ethical challenges, including bias, privacy concerns, and the potential for job displacement. Leadership prioritizes responsible AI practices and proactively addresses these issues to ensure ethical development and deployment.