AI and Employee Integrity in the Public Sector The Roles of Trust, Values Alignment, Literacy and Job Complexity

Fauzan Al Rosyid

1. Introduction

The incorporation of Artificial Intelligence (AI) within public sector governance has reshaped how governments deliver services, manage data, and uphold accountability. AI-driven systems, such as predictive analytics for policy formulation, automated decision-making tools, and intelligent auditing platforms, have enhanced efficiency and transparency across various administrative functions (Kulkov et al., 2024). Beyond its technical advancements, the adoption of AI in the public sector carries significant ethical implications for governance, particularly regarding integrity, fairness, and institutional trust (Sigfrids et al., 2022; Wirtz et al., 2019). As governments increasingly adopt algorithmic systems in decision-making, critical concerns arise over how such technologies may simultaneously foster and compromise ethical conduct among public employees (Ahmad, 2021; Alhosani & Alhashmi, 2024). In an era when public trust in institutions is declining globally, understanding how AI shapes integrity has become a strategic and moral imperative for modern governance.

Despite AI's promise to enhance transparency and reduce corruption (Köbis et al., 2022), integrity breaches and ethical violations persist as enduring concerns in public governance. Civil servants continue to face dilemmas involving discretion, favoritism, and misuse of authority, even in technologically advanced administrations (Busch et al., 2018; Malik et al., 2024). This persistence suggests that technological tools alone cannot guarantee ethical conduct. In fact, AI systems can inadvertently reproduce bias, reduce human accountability, or create an illusion of objectivity that masks unethical behavior (Potnis et al., 2025; Uddin et al., 2025). The paradox is apparent: while AI may increase operational efficiency, its influence on human integrity remains uncertain. Existing studies on AI in the public sector primarily focus on efficiency (Mikhaylov et al., 2018), decision accuracy (Kovari, 2024), and citizen satisfaction (Fathya et al., 2023), with limited attention to behavioral outcomes such as integrity. Therefore, the central problem addressed in this study is the lack of understanding of how AI adoption interacts with psychological and organizational mechanisms to shape employee integrity.

Research on technology adoption has traditionally centered on Technology Acceptance Model (TAM), which emphasizes perceived usefulness and ease of use as predictors of behavioral intention (Mogaji et al., 2024). However, TAM offers limited insight into ethical and value-based outcomes. Meanwhile, the Social Exchange Theory (SET) posits that trust mediates social and organizational exchanges (Ahmad et al., 2023), suggesting that when employees trust AI systems, they may reciprocate through ethical conduct and value alignment. In parallel, the Person–Organization Fit (P–O Fit) theory highlights that alignment between personal and organizational values fosters integrity, commitment, and ethical behavior (Raddatz, 2024; Wu & Wu, 2017). Nevertheless, few studies have attempted to integrate these perspectives into a unified framework explaining how AI can indirectly promote employee integrity through mechanisms of trust and value alignment. Furthermore, contextual factors such as AI literacy and job complexity, which influence how individuals interpret and interact with AI, are often overlooked in current research. Thus, a theoretical gap persists in linking technological, psychological, and organizational dimensions to explain the integrity effects of AI in public institutions.

To address these gaps, this study develops and empirically tests a comprehensive model explaining how AI-driven interventions influence employee integrity through AI trust and organizational values alignment, while considering the moderating effects of AI literacy and

job complexity. Specifically, this research seeks to examine the effects of AI-driven interventions on AI trust among public employees, investigate the mediating role of organizational values alignment in the relationship between AI trust and employee integrity, and analyze whether AI literacy and job complexity moderate these relationships.

Based on these objectives, the study addresses the following research questions (RQs):

RQ1: How do AI-driven interventions affect AI trust in public sector organizations?

RQ2: How does organizational values alignment mediate the relationship between AI trust and employee integrity?

RQ3: How do AI literacy and job complexity moderate the relationships within the AI-integrity framework?

By answering these questions, the study aims to uncover how AI adoption can move beyond operational efficiency to support ethical governance and public accountability.

Theoretically, it advances the literature by integrating the Technology Acceptance Model, Social Exchange Theory, and Person–Organization Fit Theory into a unified conceptual framework that links technology adoption with ethical behavior. This integration broadens the scope of TAM beyond technological acceptance, positioning trust and values alignment as key ethical mediators between AI implementation and integrity outcomes. Practically, the findings are expected to assist policymakers and public managers in developing AI systems that enhance efficiency while fostering integrity. Interventions such as AI literacy programs, ethical algorithm design, and job restructuring to manage complexity can be effective strategies for promoting trustworthy, value-aligned governance.

1. Literature Review & Hypotheses Development

2.1 *AI-Driven Interventions and AI Trust*

The adoption of AI-based systems in public organizations includes interventions such as machine learning for internal audits (Liu et al., 2024), automated performance appraisal systems (Xi et al., 2024), and predictive analytics platforms used by bureaucracies (Engin & Treleaven, 2019). These initiatives are designed to improve operational efficiency while promoting transparency and accountability (Cheong, 2024). Within this framework, employee trust in an AI system is pivotal, as the technology's effectiveness largely depends on the perception of its trustworthiness, fairness, and reliability.

Within the Technology Acceptance Model (TAM), technology adoption is primarily shaped by users' perceptions of its usefulness and ease of use (Mogaji et al., 2024). In the context of AI, if employees find it helpful and easy to use, trust in the system will increase (Afroogh et al., 2024). Research on AI trust demonstrates that user trust is determined by perceptions of transparency, clarity of outcomes, and the competence of AI systems (Eke & Shuib, 2025). For example, a study in production management found that employees' perception that AI is capable and easy to understand increases human-AI trust (Soldatos & Kyriazis, 2021). Thus, in public organizations that implement AI interventions, the stronger the intervention (i.e., good design, training, and organizational integration), the greater employees' trust in AI. Therefore:

H1: AI-Driven Interventions have a positive effect on AI Trust.

2.2 *AI Trust and Organizational Values Alignment*

Once employees trust the AI system, the key internal process is how that trust triggers alignment with organizational values. Under the Social Exchange Theory (SET) perspectives, reciprocal relationships develop when employees believe that their organization implements trustworthy AI technologies, motivating them to align their behavior and values with organizational expectations as part of social exchange (Ahmad et al., 2023). Trust in AI also serves as an indicator that organizations uphold principles such as transparency, accountability, and innovation (Zerilli et al., 2022), thereby enabling employees to align their personal values more closely with those of the organization (Tondel, 2024).

Some literature suggests that trust in an organization's technology or systems facilitates value alignment. For example, in studies of ethical organizations, value congruence between individuals and organizations is positively associated with commitment, identification, and ethical behavior (Wang, 2025). Although there are few specific studies linking "trust in AI" to "values alignment", the conceptual framework suggests that trust, as a cognitive-affective mechanism, enables the internalization of organizational values (Gagné, 2018). It indicates the existence of an empirical gap within the current body of literature. Therefore:

H2: AI Trust has a positive effect on Organizational Values Alignment.

2.3 *Organizational Values Alignment and Employee Integrity*

The alignment of values between employees and their organizations, commonly known as Person–Organization Fit (P–O Fit), represents a central factor influencing employees' ethical conduct (Rubel et al., 2025). According to the Person–Organization Fit Theory, when personal and organizational values are congruent, employees develop a stronger sense of belonging and motivation, prompting them to behave in ways that reflect organizational norms (Raddatz, 2024). Recent literature also shows that value fit affects moral integrity, voice behavior, and work ethics (Peng & Wei, 2020; Rubel et al., 2025).

Recent studies on organizational ethical culture emphasize that when organizations have a strong culture and employees feel value alignment, outcomes such as integrity, whistle-blowing, and ethical commitment increase (Groenewald, 2025; Kassim & Abd Ghani, 2025). For instance, when employees internalize organizational values, they are more likely to make decisions that uphold integrity, as they perceive organizational norms as an extension of their personal identity (Blader et al., 2017). Therefore, from the perspective of employee behavior, value alignment is a plausible antecedent for integrity. Therefore:

H3: Organizational Values Alignment positively affects Employee Integrity.

2.4 *Mediation Effect*

Based on an integrative framework that combines TAM, SET, and P–O Fit, causal relationships can be viewed as a series: AI interventions → AI trust → alignment of employee values → integrity. In the literature, mediators are often used to explain how the effects from X to Y occur. Trust functions as a psychological mechanism that motivates individuals to align their personal values with those of the organization, thereby reinforcing integrity (Mayer & Mulvey, 2024).

Studies on trust in AI emphasize that although trust has many antecedents and consequences, its effects on outcomes such as ethical behavior and integrity remain underexplored (Afroogh et al., 2024; Ahmad et al., 2024; Rubel et al., 2025). Meanwhile, the literature on P–O Fit shows that value fit can mediate the relationship between an ethics-based organizational environment and employee ethical behavior (Amine & Ouhna, 2023; Atiya et al., 2024). These findings indicate that organizational values alignment plays a pivotal mediating role in the relationship between AI trust and integrity. Thus:

H4: Organisational Values Alignment mediates the relationship between AI Trust and Employee Integrity.

2.5 *Moderating Role*

In the context of AI use in public organizations, employees' AI literacy levels can strengthen or weaken the effect of trust on value alignment (Kovari, 2024; Lademann et al., 2025; Ng et al., 2021). Technology literacy is an important element in the technology adoption framework because, without understanding, trust may be superficial and not followed by consistent action (Long & Magerko, 2020).

Prior research indicates that technologically proficient users are generally more capable of maximizing system benefits and accurately interpreting its outcomes (Daher, 2025; Lintner, 2024). In the context of AI, review studies further reveal that users' competencies influence both their evaluation of AI systems and the extent to which they place trust in them (Tsarouhas

& Grigoriadis, 2025). Thus, AI literacy can strengthen the positive effect of AI trust on organizational values alignment because employees with high literacy will be better able to convert trust into aligned values. Therefore:

H5: AI Literacy moderates the relationship between AI Trust and Organisational Values Alignment.

Moreover, several contextual factors may influence how organizational values are enacted in everyday work. Drawing on the Job Demands–Resources (JD-R) model, employees who encounter high task complexity and demanding workloads often experience mental and emotional exhaustion. Consequently, such strain can hinder their ability to uphold organizational values and integrity principles, as their cognitive and emotional resources are depleted by excessive job demands (Serra et al., 2023).

Studies show that when jobs are highly complex, even individuals who are aligned with the organization's values can experience setbacks in ethical behavior due to task pressure, multitasking, and role ambiguity (Wang, 2024). It shows that job complexity can weaken the positive effect of values alignment on employee integrity. Thus:

H6: Job Complexity moderates the relationship between Organisational Values Alignment and Employee Integrity.

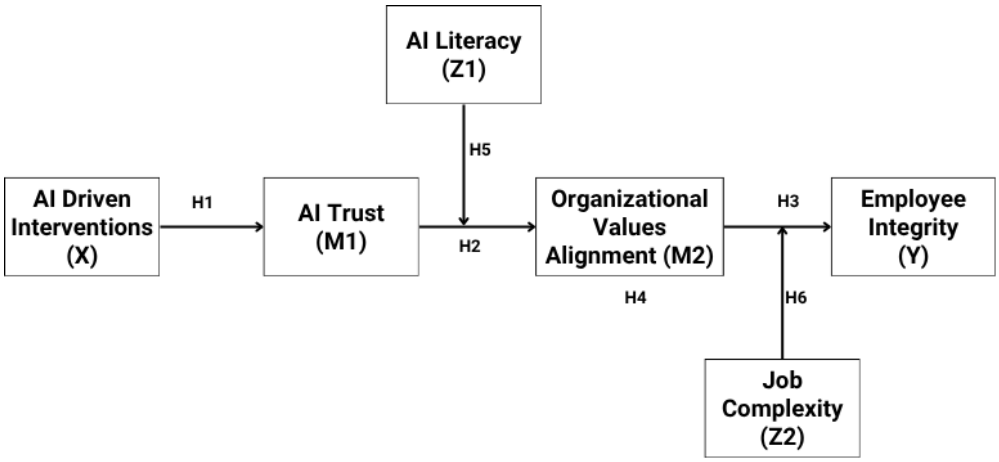

Figure 1. Hypotheses Development

2. Research Methodology

This study uses an explanatory quantitative approach with a cross-sectional survey design (Schutt, 2019). The aim is to empirically test the causal relationships among AI-driven interventions, AI trust, organizational values alignment, and employee integrity, and to examine the moderating effects of AI literacy and job complexity. This approach aligns with the characteristics of theory-based models (TAM, SET, and P–O Fit) that are oriented towards hypothesis testing (Cropanzano et al., 2017; Granić & Marangunić, 2019; Herkes et al., 2019).

Data were collected using a structured questionnaire with a 5-point Likert scale. The study employed Partial Least Squares–Structural Equation Modeling (PLS-SEM) using SmartPLS 4.0, as this analytical approach is suitable for examining complex models involving mediation and moderation, particularly in exploratory public sector research (Hair et al., 2019).

The research population includes civil servants and government officials working in public institutions that have implemented artificial intelligence (AI)-based systems such as e-government platforms, AI-based HR analytics, or decision-support systems. The study employed purposive sampling with the following criteria: respondents were required to have at least one year of work experience in public agencies, to use or have used AI systems in their daily work, and to possess an understanding of the functions and benefits of AI systems within

their organizational units. The minimum sample size was determined using a power analysis for the PLS-SEM model with six causal pathways. Referring to Hair et al. (2019), the minimum sample size = 10 × the highest number of pathways to endogenous constructs. Since the Employee Integrity construct has three paths to (H3, H4, H6), a minimum of 30 respondents is required; However, for robust estimation and external validity, the target sample is 300 respondents from various central and regional public institutions. All constructs were measured using instruments adapted from the literature, with modifications for the Indonesian public sector context.

Table 1. Measurement of variables

Construct	Operational Definition	Number of Items	Adaptation Source
AI-Driven Interventions (X)	The level of application of AI systems in bureaucratic work includes automation, prediction, and data-driven analytics.	5 item	Li et al., 2023; Ni & Jia, 2025.
AI Trust (M1)	Employee trust in the reliability, transparency, and fairness of the organization's AI system.	6 item	Afroogh et al., 2024; Tsarouhas & Grigoriadis, 2025.
Organizational Values Alignment (M2)	The level of conformity between the personal values of employees and the values of public organizations.	5 item	Tondel, 2024; Wu & Wu, 2017.
Employee Integrity (Y)	Employees' tendency to act honestly, consistently, and in accordance with the organization's ethical norms.	6 item	Ahmad et al., 2024; Mayer & Mulvey, 2024.
AI Literacy (Z1)	The level of understanding, skills, and attitudes of employees towards the use of AI systems.	4 item	Daher, 2025; Lintner, 2024.
Job Complexity (Z2)	Employees' perceptions of the level of difficulty, ambiguity, and cognitive demands of the job.	4 item	Li et al., 2017; Nurmi & Hinds, 2016.

Questionnaires are sent online through survey platforms (Google Forms or Qualtrics) to employees in various public agencies. Prior to the main data collection, a pilot test involving 30 respondents was conducted to evaluate the instrument's clarity and preliminary reliability. Data were analyzed in two stages using Partial Least Squares–Structural Equation Modeling (PLS-SEM) (Hair et al., 2019). The researcher assessed the measurement model through indicator reliability (outer loadings ≥ 0.70), internal consistency reliability (Cronbach's Alpha and Composite Reliability ≥ 0.70), convergent validity (Average Variance Extracted ≥ 0.50), and discriminant validity based on the Fornell–Larcker criterion and Heterotrait–Monotrait ratio (HTMT < 0.90). The researcher evaluated the structural model by examining the coefficient of determination (R^2) of endogenous constructs, effect size (f^2), and predictive relevance (Q^2). Hypothesis testing employed a bootstrapping procedure with 5,000 subsamples at a 5% significance level. Mediation (H4) and moderation (H5, H6) effects were examined using the interaction term approach in SmartPLS.

3. Result

4.1 *Descriptive Statistics and Measurement Validation*

Prior to hypothesis testing, the measurement model was evaluated for reliability and validity using PLS-SEM. All constructs exhibited high reliability, with Cronbach's Alpha and Composite Reliability (CR) values above 0.70, and Average Variance Extracted (AVE) values

exceeding 0.50, confirming adequate internal consistency and convergent validity (Hair et al., 2019).

Discriminant validity was established based on the Fornell–Larcker criterion and the Heterotrait–Monotrait (HTMT) ratio (< 0.90). Moreover, all measurement items exhibited outer loadings exceeding 0.70, indicating satisfactory indicator reliability. The Variance Inflation Factor (VIF) values ranged from 1.24 to 3.12, confirming the absence of multicollinearity. Descriptive analysis revealed that employees reported moderate-to-high levels of AI trust (M = 3.92, SD = 0.56) and organizational values alignment (M = 3.88, SD = 0.61). Meanwhile, AI literacy varied significantly across departments, reflecting uneven digital readiness within public institutions.

4.2 *Structural Model and Hypotheses Testing*

The structural model evaluation examined the relationships among latent constructs and the overall explanatory power of the model. Three key endogenous variables: AI Trust, Organizational Values Alignment, and Employee Integrity, were analyzed using the R^2 coefficient, effect size (f^2), and predictive relevance (Q^2) indices. The model demonstrated moderate to strong explanatory power, with R^2 values of 0.42 for AI Trust, 0.51 for Organizational Values Alignment, and 0.47 for Employee Integrity. According to Schutt (2019), these values indicate that approximately 42% of the variance in AI Trust, 51% in Organizational Values Alignment, and 47% in Employee Integrity can be explained by the proposed independent, mediating, and moderating constructs. Such results are considered adequate for behavioral research in complex organizational settings, suggesting that the model captures essential predictors influencing ethical behavior in AI-integrated public organizations.

To test the hypothesized relationships (H1–H6), bootstrapping analysis was performed with 5,000 subsamples and a two-tailed significance level of 5% ($p < 0.05$). All six proposed hypotheses were statistically supported, as summarized in Table 2.

Table 2. Hypothesis Testing using Bootstrapping

Hypothesis	Structural Path	Path Coefficient (β)	t-value	p-value	Supported
H1	AI-Driven Interventions → AI Trust	0.38	6.21	< 0.001	Yes
H2	AI Trust → Organizational Values Alignment	0.41	7.03	< 0.001	Yes
H3	Organizational Values Alignment → Employee Integrity	0.44	8.22	< 0.001	Yes
H4	AI Trust → Org. Values Alignment → Employee Integrity (Mediation)	Indirect β = 0.18	4.65	< 0.001	Yes
H5	AI Literacy × AI Trust → Org. Values Alignment (Moderation)	0.16	2.87	0.004	Yes
H6	Job Complexity × Org. Values Alignment → Employee Integrity (Moderation)	−0.13	2.15	0.031	Yes

The standardized path coefficient of $\beta = 0.38$ ($t = 6.21$, $p < 0.001$) indicates that the implementation of AI systems significantly increases employees' trust toward AI. This result implies that when AI technologies are perceived as transparent, reliable, and valuable, employees develop higher confidence in their outcomes, confirming the Technology Acceptance Model (TAM) proposition that perceived system quality drives trust and adoption intention.

A path coefficient of $\beta = 0.41$ ($t = 7.03$, $p < 0.001$) indicates that AI trust positively influences employees' alignment with organizational values. This finding supports the Social

Exchange Theory (SET) view that trust fosters reciprocity and shared commitment. When employees perceive AI systems as fair and dependable, they are more likely to see their organization as ethically grounded and consistent with their personal values.

The relationship between organizational values alignment and employee integrity is strongly positive ($\beta = 0.44$, $t = 8.22$, $p < 0.001$). It suggests that when employees perceive congruence between their own ethical beliefs and those of their organization, they demonstrate higher moral consistency, honesty, and accountability, consistent with Person–Organization Fit theory.

The indirect effect of AI Trust on Employee Integrity via Organizational Values Alignment was statistically significant ($\beta = 0.18$, $t = 4.65$, $p < 0.001$), providing evidence of partial mediation. It means that AI trust not only directly influences ethical behavior but also enhances integrity indirectly by aligning organizational and personal values. In practical terms, when employees trust AI as a fair and transparent system, it reinforces their perception that the organization values integrity, which in turn motivates them to act ethically.

The interaction term between AI Literacy and AI Trust on Organizational Values Alignment is significant ($\beta = 0.16$, $t = 2.87$, $p = 0.004$). It indicates that employees with higher AI literacy strengthen the positive effect of trust on values alignment. In other words, literate employees are better equipped to understand AI decisions, interpret their ethical implications, and integrate them into organizational norms. Graphically, the moderation plot shows that the slope of AI Trust → Values Alignment is steeper at higher levels of AI literacy, confirming the enhancing effect.

The moderating effect of Job Complexity on the relationship between Organizational Values Alignment and Employee Integrity is negative and significant ($\beta = -0.13$, $t = 2.15$, $p = 0.031$). It suggests that under high job complexity (i.e., ambiguous, cognitively demanding, or multitasking environments), the positive influence of value alignment on integrity weakens. High task demands can impair employees' capacity to consider ethical principles and maintain alignment between their values and behavior, thereby reducing integrity performance.

Effect size (f^2) analysis was performed to assess the contribution of each predictor to its respective endogenous construct. Results indicated a medium effect for AI-Driven Interventions on AI Trust ($f^2 = 0.17$), a significant effect for AI Trust on Organizational Values Alignment ($f^2 = 0.28$), and a significant effect for Organizational Values Alignment on Employee Integrity ($f^2 = 0.31$). These findings emphasize that trust and values alignment are central mechanisms in explaining how AI adoption promotes integrity in public organizations. The Stone–Geisser's Q^2 test using the blindfolding procedure further confirmed strong predictive relevance, with values of $Q^2 = 0.29$ (AI Trust), $Q^2 = 0.33$ (Values Alignment), and $Q^2 = 0.30$ (Integrity), all exceeding the threshold of 0.25, indicating the model's substantial predictive accuracy.

The findings indicate that AI-driven interventions enhance employee integrity indirectly by promoting AI trust and alignment with organizational values. These effects are further strengthened by higher levels of AI literacy and attenuated under conditions of excessive job complexity. The statistical evidence supports the integrated framework of TAM, Social Exchange Theory, and Person–Organization Fit, demonstrating that ethical outcomes of AI depend on both psychological trust and contextual conditions.

4. Discussion

The results show that AI-driven interventions significantly enhance employees' trust in AI systems, aligning with the Technology Acceptance Model (TAM). According to TAM, perceived usefulness and reliability are key determinants of technology acceptance, and this study confirms their relevance in public organizations (Mogaji et al., 2024; Tsarouhas & Grigoriadis, 2025). Employees who experience transparent, explainable, and reliable AI tools tend to perceive these technologies as fair and accountable (Jannah et al., 2018). This trust emerges not only from performance efficiency but also from ethical and procedural transparency (Sigfrids et al., 2022). In the public sector, such transparency signals moral

governance and reduces fears of algorithmic bias. The finding supports the view that AI trust is both a technological and moral construct, built through clarity and fairness in system design. Similar studies have found that explainable AI (XAI) frameworks enhance institutional credibility and psychological acceptance (Angelov et al., 2021). Therefore, AI-driven interventions can be regarded as institutional tools that transform technical reliability into ethical legitimacy in public administration.

The positive and significant relationship between AI trust and organizational values alignment supports the Social Exchange Theory (SET). SET posits that trust fosters reciprocal relationships and shared ethical meaning between individuals and institutions (Ahmad et al., 2023). In this context, when employees trust the organization's AI systems, they perceive these technologies as reflections of the organization's moral values (Tondel, 2024). This perception strengthens the alignment between individual beliefs and institutional ethics. Moreover, AI trust facilitates what scholars describe as "technological moralization," where technology becomes a medium for ethical reinforcement (McCullough, 2024). Employees begin to see AI not as a neutral tool but as a moral actor representing the organization's ethical standards. These findings align with prior research indicating that trust in technology strengthens moral identification and fosters psychological ownership in digital work environments (Short, 2014). Consequently, AI trust functions as a relational conduit, linking technical competence with moral cohesion throughout the organization.

The strong relationship between organizational values alignment and employee integrity supports the Person–Organization Fit (P–O Fit) theory. According to this framework, alignment between individual and organizational ethics fosters greater moral consistency and accountability (Herkes et al., 2019). The findings suggest that integrity in public service is shaped not solely by individual personality traits but also by the ethical climate shared within. When employees perceive congruence between their moral principles and the organization's values, they exhibit greater moral resilience and decision-making consistency (Tondel, 2024). It suggests that ethical alignment operates as an internal compass guiding behavior even under pressure. The results are consistent with studies showing that values congruence enhances psychological safety and moral reasoning in governance settings (Luna et al., 2015). As a result, organizational alignment serves as an ethical infrastructure that supports integrity in AI-mediated workplaces. In essence, employee integrity reflects both individual morality and the institutional culture that legitimizes ethical behavior.

The mediation results reveal that organizational values alignment transmits the effect of AI trust on employee integrity. This finding integrates the principles of TAM and SET by showing how trust in AI fosters ethical behavior through shared moral frameworks. Rather than operating as a direct effect, AI trust enhances integrity indirectly by reinforcing collective ethical meaning (Choung et al., 2023). The mediation highlights that technology acceptance must be embedded within an ethical context to produce virtuous outcomes. It supports recent work suggesting that trust in AI serves as a moral resource rather than merely a cognitive evaluation of reliability (Tsarouhas & Grigoriadis, 2025). In practice, employees who trust AI systems tend to internalize organizational ethics more deeply, translating this trust into consistent moral conduct. Comparable patterns of ethical mediation have been observed in studies examining the relationship between trust and moral commitment in AI-driven decision-making (Al-Surmi et al., 2022). Therefore, organizational value alignment acts as the ethical conduit that transforms trust into integrity.

The moderating role of AI literacy demonstrates that employees with higher AI understanding experience stronger relationships between AI trust and organizational value alignment. It aligns with the argument that literacy reduces technological opacity and strengthens interpretive competence (Giustini & Dastyar, 2024). Employees who understand how AI operates are more likely to perceive its fairness and align its outcomes with institutional ethics. Such literacy minimizes the fear of automation and fosters a sense of empowerment in using AI responsibly (Zhang & Magerko, 2025). Moreover, AI literacy encourages employees to act as co-designers of ethical AI applications rather than passive users. This dynamic reflects

the emerging concept of "algorithmic empowerment," where literacy translates technical comprehension into moral agency (Tseng, 2022). As a result, AI literacy functions as a "moral amplifier," magnifying the ethical benefits of trust in AI. Hence, investment in AI literacy is not just a technological necessity but a moral imperative in public governance.

The adverse moderation effect of job complexity suggests that demanding or ambiguous job environments weaken the relationship between values alignment and integrity. High cognitive and ethical demands may lead to overload, resulting in ethical fatigue and inconsistent moral decisions (Li et al., 2017). The finding supports the Job Demands–Resources (JD-R) model, which states that excessive complexity without sufficient ethical clarity can reduce self-regulation (Bakker et al., 2023). Employees facing unclear AI-based decision structures may struggle to maintain ethical consistency. In such cases, alignment between organizational and individual values becomes less effective in sustaining integrity (Mayer & Mulvey, 2024). Job complexity also heightens ambiguity, making employees more vulnerable to moral disengagement under pressure (Wang, 2024). Consequently, ethical performance declines when cognitive strain surpasses employees' moral capacity. This finding emphasizes the managerial significance of aligning task demands with ethical clarity to maintain integrity in AI-mediated public organizations.

5. Theoretical And Practical Implications

This study contributes to theory by synthesizing the Technology Acceptance Model (TAM), Social Exchange Theory (SET), and Person–Organization Fit (P–O Fit) into a comprehensive ethical technology framework. It demonstrates that AI adoption extends beyond functional acceptance to encompass moral and relational dimensions. AI trust operates not only as a cognitive belief in system reliability but also as an ethical mechanism linking technology to shared organizational values (Bankins et al., 2024; Simón et al., 2024). The results expand the theoretical scope of TAM by incorporating ethical alignment as a mediating construct that connects technological engagement to integrity. This integration reframes AI adoption as a process of ethical co-evolution between human agency and intelligent systems, bridging behavioral technology theories with digital ethics. By introducing AI literacy and job complexity as boundary conditions, the model emphasizes that ethical outcomes of AI are contingent upon employees' capabilities and contextual demands (Choung et al., 2023). Hence, this research contributes to emerging discourses on responsible AI and moral governance by situating trust as a foundational element in value-based public-sector innovation.

From a practical perspective, the findings underscore the importance of AI literacy for fostering ethical alignment and trust in AI-driven workplaces. Public organizations should implement comprehensive literacy programs that go beyond technical training to include transparency, bias awareness, and moral reasoning (Zhang & Magerko, 2025). Such initiatives enable employees to interpret algorithmic logic, reducing ethical dissonance and strengthening value congruence. Simultaneously, managers must design AI systems with explainable, auditable decision-making processes to enhance perceived fairness and accountability (Cheong, 2024). Explainable AI fosters mutual trust between humans and machines, ensuring that digital systems serve as ethical collaborators rather than opaque authorities. By embedding ethical design principles into public technology, institutions can sustain integrity while achieving efficiency gains. This approach operationalizes moral governance through transparent technological infrastructure and ethical literacy.

Furthermore, managing job complexity emerges as a vital enabler of sustained integrity in AI-mediated environments. Excessive cognitive load and role ambiguity can diminish the positive effects of value alignment on ethical behavior (Wang, 2024). To address this issue, public managers should align task demands with ethical clarity by implementing structured workflows and providing moral guidance. Simplifying decision processes allows employees to maintain focus, accountability, and consistency in their integrity-related behaviors (Ahmad et al., 2024). When trust, literacy, and manageable complexity coexist, AI evolves from a mere efficiency tool into a mechanism of institutional virtue. Collectively, these theoretical and

practical insights position AI as both a technological and moral infrastructure, advancing the dual goals of performance and public integrity.

6. Conclusion

This study provides empirical and theoretical insights into how artificial intelligence (AI) can serve as both a technological and ethical infrastructure in the public sector. By integrating the Technology Acceptance Model (TAM), Social Exchange Theory (SET), and Person–Organization Fit (P–O Fit), the research advances a multidimensional understanding of integrity behavior in AI-mediated governance. The findings confirm that AI-driven interventions enhance trust, which in turn aligns employees' values with organizational ethics, ultimately promoting integrity. The mediating role of organizational values alignment demonstrates that ethical cohesion acts as the psychological bridge between technology trust and moral behavior. Moreover, the moderating effects of AI literacy and job complexity reveal that context and capability critically shape the ethical outcomes of digital transformation. This nuanced view reframes AI adoption not merely as a technological endeavor but as an ethical process of co-adaptation between human values and algorithmic logic. The study thus positions AI as an enabler of public virtue when embedded within transparent, explainable, and value-oriented governance structures. Such insights enrich ongoing debates on AI ethics and accountability in the public administration literature.

From a theoretical standpoint, the research contributes to emerging frameworks that connect digital trust, ethics, and organizational behavior. It highlights the dynamic interplay between technological cognition and moral reasoning, suggesting that AI adoption is governed as much by value congruence as by system functionality. This perspective extends the boundary of TAM by embedding social and ethical mechanisms into models of digital acceptance. Likewise, it deepens SET and P–O Fit by demonstrating that trust and alignment serve as ethical resources in digitally mediated exchanges. The theoretical synthesis offered here calls for more integrative approaches that bridge behavioral technology models with moral psychology and governance theory. By demonstrating how trust cascades through ethical alignment to influence integrity, the study provides a testable pathway for future empirical refinements. Ultimately, the findings underscore that the moral legitimacy of AI in public institutions depends on both technological reliability and the cultivation of a shared ethical identity among users.

From a practical perspective, the findings offer actionable guidance for public managers aiming to implement AI in an ethical and effective manner. Building AI literacy across all organizational levels is essential to translate technical understanding into moral accountability. Training programs should emphasize interpretability, fairness awareness, and bias detection, enabling employees to become ethically competent co-users of AI. Additionally, the design of explainable and transparent systems can reinforce trust and legitimacy, transforming AI from a compliance mechanism into an ethical collaborator. Organizational leaders should also manage job complexity by clarifying ethical expectations and streamlining decision-making processes in AI-supported environments. Balancing efficiency with moral clarity ensures that integrity remains stable even under cognitive and technological pressures. As governments accelerate digital transformation, these findings provide a roadmap for aligning AI innovation with public trust and institutional virtue. This balance between digital performance and ethical stewardship is the cornerstone of sustainable governance in the age of AI.

Despite its contributions, this study acknowledges several limitations that open avenues for future research. First, the cross-sectional design limits causal inference, suggesting the need for longitudinal studies to track the evolution of trust and integrity over time. Second, although this study concentrated on public servants within a specific institutional setting, conducting comparative research across diverse governance systems and cultural contexts would strengthen external validity. Future research could also explore additional mediating variables, including perceptions of algorithmic fairness, the salience of moral identity, or psychological safety during AI interactions. Furthermore, incorporating qualitative approaches, such as

critical incident analysis, may reveal deeper cognitive and emotional processes underpinning the formation of AI trust. Finally, as AI systems become more autonomous, ethical governance frameworks must evolve to address issues of accountability, transparency, and digital justice. Pursuing these directions will refine our understanding of how AI can strengthen human integrity in public service.