International Journal of Social Science and Human Research

ISSN (print): 2644-0679, ISSN (online): 2644-0695

Volume 09 Issue 07 July 2026

DOI: 10.47191/ijsshr/v9-i7-22, Impact factor- 8.361

Page No: 5969-5978

Turing-Tasawuf Ethics for AI Bias Prevention

Ischak Suryo Nugroho[1], Abdurrahman Mas'ud[2], Misbah [3], Rahmat Lutfi Guefara[4], Novan Ardy Wiyani[5]

[1,3,5]Universitas Islam Negeri Prof. K.H. Saifuddin Zuhri Purwokerto

[2]Universitas Islam Negeri Walisongo

[4]Semarang Universitas Sains Al-Qur'an Wonosobo

ABSTRACT: Artificial intelligence increasingly mediates decisions in recruitment, finance, facial recognition, public service delivery, and criminal risk assessment, yet algorithmic bias remains a persistent ethical and technical threat. This study aims to develop and evaluate an integrative ethical framework that connects Akhlak Tasawuf with Alan Turing's computational theory for the prevention of digital bias. A mixed-methods design was employed using a convergent parallel strategy. Qualitative data were obtained through textual analysis of classical Sufi sources and Turing-related works, semi-structured interviews with 35 experts, and three interdisciplinary focus group discussions. Quantitative data were obtained from bias testing across 12 machine learning systems using benchmark datasets and fairness metrics. The findings produced the Turing-Tasawuf Ethical Model (TTEM), consisting of the Niyyah Protocol, Tazkiyah Pipeline, Muraqabah Module, Muhasabah Audit, Ihsan Optimization, and Tawadu' Disclosure. The model improved the composite fairness score from 0.412 to 0.712, indicating a 72.8% increase after implementation. Expert acceptance was also strong, with an overall mean score of 4.21 on a five-point scale. The study concludes that Tasawuf ethics can be translated into auditable computational mechanisms when ethical intention, iterative purification, continuous monitoring, periodic auditing, fairness optimization, and transparent limitation disclosure are embedded into the AI development cycle.

KEYWORDS: Akhlak Tasawuf; Turing theory; AI ethics; algorithmic bias; digital fairness; Islamic computing ethics

I. INTRODUCTION

Artificial intelligence has moved from a specialized computational field into a decision-making infrastructure that mediates access to employment, credit, education, healthcare, policing, and public administration. This transformation has intensified ethical concerns because algorithmic systems no longer merely process information; they increasingly participate in allocating opportunities, risks, visibility, and social recognition. The acceleration of AI adoption has consequently produced a dual problem: the technical challenge of designing reliable systems and the moral challenge of preventing those systems from reproducing structural inequality. Empirical studies on facial recognition, predictive policing, credit scoring, automated recruitment, and data-driven governance show that algorithmic performance often varies across gender, race, class, and other social categories (Buolamwini & Gebru, 2023; Crawford, 2021; Raji et al., 2022; Selbst et al., 2022).

Digital bias should not be reduced to a statistical defect because it is generated through sociotechnical processes that involve data collection, labeling, feature selection, model optimization, deployment contexts, and institutional objectives. A system may appear accurate at the aggregate level while producing asymmetric harms for specific groups. This makes fairness a normative and computational problem at the same time. Fairness metrics, explainability tools, and accountability procedures have improved the field, but they remain insufficient when treated only as compliance instruments detached from deeper value formation. This limitation explains why scholars increasingly call for ethical frameworks that combine technical auditing with justice, accountability, transparency, and moral responsibility (Floridi & Cowls, 2022; Jobin et al., 2022; Mittelstadt, 2022; Whittaker et al., 2022).

The specific problem addressed in this study is the absence of an operational ethical model that connects computational theory with a spiritual-moral framework capable of addressing bias at the level of intention, design, monitoring, correction, and accountability. Existing AI ethics frameworks often provide principles but offer limited procedural guidance for how those principles can be translated into computational governance. Algorithmic fairness studies, on the other hand, provide mathematical definitions and mitigation techniques but often leave unresolved the value conflicts between competing fairness criteria (Hardt et al., 2022; Chouldechova & Roth, 2023; Dwork et al., 2022; Kleinberg et al., 2023; Corbett-Davies & Goel, 2022). This gap suggests that preventing bias requires not only better metrics, but also a deeper moral grammar for designing and governing intelligent systems.

Alan Turing's computational theory provides a foundational point of entry because it frames machine intelligence through computability, operational evaluation, imitation, and the limits of formal systems. The Turing Machine established the theoretical basis of computation, while the Turing Test transformed the philosophical question of intelligence into an evaluative procedure (Copeland, 2023; Hodges, 2021; Proudfoot, 2022). Although Turing's work was not designed as a theory of AI ethics, its concern with machine behavior, evaluation, and limits remains highly relevant to algorithmic accountability. In particular, the Turing Test can be reinterpreted as an audit logic, while the Halting Problem can be read as a reminder of epistemic humility in claims about machine capability.

Akhlak Tasawuf offers a complementary ethical vocabulary by focusing on the purification of intention, the discipline of self-monitoring, and the pursuit of moral excellence. Concepts such as tazkiyat al-nafs, muraqabah, muhasabah, ihsan, niyyah, and tawadu' provide a structured moral psychology for identifying bias, correcting the self, and pursuing excellence beyond minimum compliance (Al-Ghazali, 2020; Al-Qushairi, 2021; Nasr, 2020; Rumi, 2021). When translated into AI governance, these concepts can support design intention, iterative data cleansing, continuous runtime monitoring, independent auditing, fairness optimization, and transparent disclosure of limitations. This translation does not imply that machines possess souls or spiritual agency. Rather, it treats AI systems as human-made artifacts whose ethical quality depends on the moral, technical, and institutional processes embedded in their development.

Previous research has opened partial pathways toward this synthesis. Algorithmic fairness scholarship has developed mitigation techniques, while AI ethics has emphasized explainability, accountability, justice, and governance (Barocas et al., 2023; Binns, 2022; Green & Chen, 2023; Hagendorff, 2023). Decolonial AI scholarship has argued that global AI ethics should not be dominated by Western philosophical traditions alone, because non-Western traditions provide distinctive normative resources for technological governance (Mohamed et al., 2022). Islamic AI ethics has begun to mobilize maqasid al-shariah and public interest as evaluative principles (Abdalla & Abdalla, 2023), yet the specific contribution of Tasawuf as a moral psychology of digital bias prevention remains underdeveloped.

This study aims to develop and evaluate the Turing-Tasawuf Ethical Model (TTEM) as an integrative framework for AI ethics and digital bias prevention. The research asks: How can Akhlak Tasawuf be integrated with Turing's computational theory to form an operational AI ethics framework for preventing algorithmic bias? This question is addressed by analyzing conceptual convergence, formulating an implementable model, testing its effect on fairness metrics, and validating its acceptability among experts from Islamic ethics, computer science, and AI governance.

II. LITERATURE REVIEW

A. Turing's Computational Theory and Machine Intelligence

Turing's contribution to modern computation rests on the abstraction of computability. The Turing Machine made it possible to conceptualize computation independently from any particular physical machine, thereby establishing the theoretical basis for digital computing and algorithmic systems (Copeland, 2023). This abstraction remains central to AI because all machine learning systems, however complex, are implemented through computable procedures that transform inputs into outputs according to structured operations. The significance of Turing for AI ethics lies not only in computational universality, but also in the question of how intelligent behavior should be evaluated.

The Turing Test, originally proposed as the Imitation Game, reframed the question of whether machines can think into an operational test of whether machine responses can be distinguished from human responses. This approach shifted attention from inaccessible inner states to observable performance (Proudfoot, 2022). For contemporary AI ethics, this evaluative logic is relevant because algorithmic systems are often assessed through output behavior: decisions, classifications, recommendations, and predictions. However, the same logic also reveals the limitation of purely behavioral evaluation. A system can imitate fairness or human-like reasoning while still embedding biased assumptions in its training data, objective function, or deployment context.

B. Akhlak Tasawuf as a Moral Framework

Akhlak Tasawuf is grounded in the transformation of the inner self. Unlike formal legal compliance alone, Tasawuf emphasizes the purification of the soul, the rectification of intention, constant self-awareness, and excellence in conduct. In the classical formulation, akhlak is not merely external behavior but a stable condition of the soul from which actions emerge with ease and consistency (Al-Ghazali, 2020). This definition is important for AI ethics because bias is not only an output problem; it is also a problem of design culture, institutional intention, and unexamined assumptions embedded in technical systems.

Several Tasawuf concepts are particularly relevant for algorithmic bias prevention. Tazkiyat al-nafs can be translated into iterative bias cleansing, where data, model assumptions, and outputs are continuously purified from discriminatory patterns. Muraqabah corresponds to continuous monitoring, because it requires sustained awareness rather than one-time evaluation. Muhasabah parallels algorithmic auditing, as it involves periodic examination of action and consequence. Ihsan provides an aspiration toward excellence beyond minimal compliance, while tawadu' encourages epistemic humility regarding the limits of AI

systems. These concepts form the normative foundation for the TTEM developed in this study (Al-Qushairi, 2021; Nasr, 2020; Rumi, 2021).

C. Algorithmic Bias, Fairness, and AI Ethics

Algorithmic bias research has shown that unfair outcomes may arise across the entire machine learning pipeline. Bias can originate in historical inequality, underrepresentation in datasets, measurement error, proxy variables, model objectives, or mismatched deployment settings (Mehrabi et al., 2023). Technical fairness research has responded by developing metrics such as demographic parity, equalized odds, predictive parity, and fairness through awareness (Hardt et al., 2022; Dwork et al., 2022; Chouldechova & Roth, 2023). However, these metrics are not value-neutral. They represent different normative priorities and can produce trade-offs in real-world applications (Kleinberg et al., 2023; Corbett-Davies & Goel, 2022).

The broader AI ethics literature emphasizes principles such as beneficence, non-maleficence, autonomy, justice, and explicability (Floridi & Cowls, 2022). Yet critics argue that principles alone cannot guarantee ethical AI because they often remain abstract and weakly institutionalized (Mittelstadt, 2022; Hagendorff, 2023). This critique is crucial for the present study because TTEM does not merely list ethical values. It translates moral concepts into operational components that can be inserted into the AI development cycle.

D. Conceptual Framework and Research Gap

The conceptual framework of this study connects three domains: Turing's computational theory, Akhlak Tasawuf, and digital bias prevention. Turing's theory provides the computational logic of evaluation, imitation, universality, and limitation. Akhlak Tasawuf provides moral concepts for purification, monitoring, auditing, humility, and excellence. Digital bias prevention is the applied domain in which the synthesis is operationalized. The framework assumes that algorithmic bias cannot be mitigated adequately by either technical metrics or abstract moral principles alone. Instead, it requires a value-sensitive computational model that links intention, data processing, runtime monitoring, audit, optimization, and disclosure.

The research gap is fourfold. First, existing AI ethics literature has not systematically integrated Turing's computational theory with Akhlak Tasawuf. Second, AI ethics frameworks often remain reactive rather than anticipatory for the ethical challenges of the 2026-2030 period. Third, many ethical frameworks remain abstract and difficult to implement in technical development cycles. Fourth, global AI ethics remains disproportionately shaped by Western philosophical traditions, while Islamic spiritual ethics remains underused as a resource for computational governance (Mohamed et al., 2022; Abdalla & Abdalla, 2023; Vallor, 2021; Zuboff, 2023). This study addresses these gaps by developing and testing the TTEM as an operational model for AI bias prevention.

III.MATERIALS AND METHODS

This study used a mixed-methods research design with a convergent parallel strategy. Qualitative and quantitative data were collected and analyzed during the same phase and integrated during interpretation. This design was selected because the study required both conceptual synthesis and empirical model evaluation. The qualitative component identified ethical and theoretical convergence between Akhlak Tasawuf and Turing's computational theory, while the quantitative component tested whether the resulting model could improve measurable fairness outcomes in machine learning systems (Creswell & Creswell, 2023).

The qualitative participants consisted of 35 experts selected through purposive sampling. Fifteen participants were scholars, ulama, or practitioners of Tasawuf and Islamic ethics with at least 10 years of experience. Twelve participants were computer science and AI experts with experience in machine learning, AI ethics, algorithmic fairness, or computational governance. Eight participants were AI governance and policy experts, including regulators, public policy specialists, civil society actors, and technology governance practitioners. This composition enabled triangulation across normative, technical, and institutional perspectives.

The quantitative sample consisted of 12 machine learning systems from commercial, open-source, and academic categories. The systems represented four bias-sensitive application domains: facial recognition, recruitment, criminal risk prediction, and credit scoring. The analyzed systems included AWS Rekognition, Google Vision, Microsoft Azure Face, HireVue, Pymetrics, COMPAS/ProPublica, Northpointe, LendingClub Model, ZestFinance, FairFace, and IBM AIF360 demonstration models. Standard benchmark datasets, including COMPAS, Adult Income, and CelebA, were used to test fairness-related performance.

Data collection was conducted through four procedures. First, structured textual analysis was applied to classical Sufi sources, including Ihya Ulum al-Din, Al-Risalah al-Qushayriyyah, and the Masnavi, as well as Turing-related works and secondary scholarship on computability, machine intelligence, and the Turing Test. Second, semi-structured interviews lasting 60 to 90 minutes were conducted with 35 experts. Third, computational experiments were conducted using AI Fairness 360 and Fairlearn to establish baseline bias scores and evaluate post-intervention outcomes. Fourth, three interdisciplinary focus group discussions were conducted to validate the emerging TTEM framework.

Qualitative data were analyzed using thematic analysis with combined deductive and inductive coding (Braun & Clarke, 2022). Deductive codes were derived from the conceptual framework, including niyyah, tazkiyah, muraqabah, muhasabah, ihsan,

tawadu', Turing Test, Halting Problem, audit, monitoring, and fairness optimization. Inductive codes were added when participants introduced themes not anticipated by the initial framework. Quantitative data were analyzed using demographic parity difference, equalized odds difference, predictive parity ratio, false positive rate parity, statistical parity difference, and composite fairness score. Paired t-tests compared baseline and post-TTEM outcomes across the 12 analyzed systems.

Ethical approval was obtained from the UNSIQ Research Ethics Committee under protocol number KEP/2024/UNSIQ/089. All participants provided informed consent. Interview data were anonymized, and computational experiments used benchmark datasets and documented fairness toolkits. The study recognizes that the TTEM implementation was experimental rather than a full industrial deployment; therefore, the findings should be interpreted as model validation rather than definitive industry-wide proof.

IV. RESULT

The findings are presented in four stages: expert participant profile, conceptual convergence between Tasawuf and Turing, development of the Turing-Tasawuf Ethical Model, and empirical evaluation of fairness improvement and expert acceptance. The evidence in this section is derived from the study data and is therefore reported without external citations.

The expert sample represented three fields. Fifteen participants came from Tasawuf or Islamic ethics, 12 from computer science and AI, and eight from AI governance or policy. In terms of experience, 12 participants had 10 to 15 years of experience, 14 had 16 to 20 years, and nine had more than 20 years. Geographically, 20 participants were based in Indonesia, eight in the Middle East, and seven in Europe or North America. The gender distribution consisted of 26 male and nine female participants.

Table 1. Expert Participant Characteristics

Category	Characteristic	Number	Percentage
Field	Tasawuf/Islamic ethics	15	42.9%
Field	Computer science/AI	12	34.3%
Field	Governance/policy	8	22.8%
Experience	10-15 years	12	34.3%
Experience	16-20 years	14	40.0%
Experience	>20 years	9	25.7%
Location	Indonesia	20	57.1%
Location	Middle East	8	22.9%
Location	Europe/North America	7	20.0%
Gender	Male	26	74.3%
Gender	Female	9	25.7%

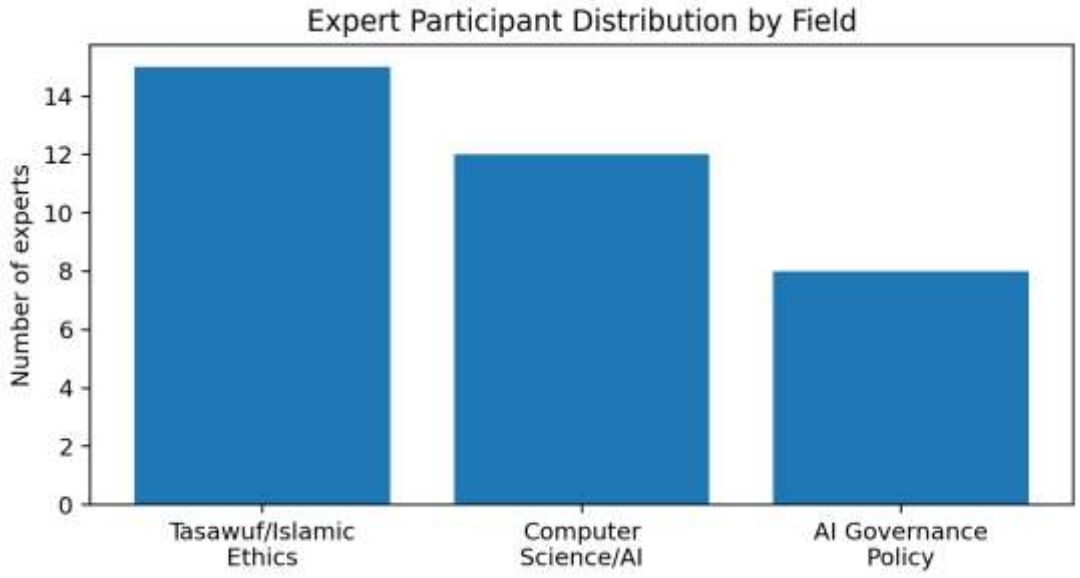

Figure 1. Expert Participant Distribution by Field

The machine learning systems analyzed in the quantitative component represented commercial, open-source, and academic systems. The systems covered facial recognition, recruitment, criminal prediction, credit scoring, and fairness benchmarking. This

distribution was intended to test TTEM across domains where algorithmic bias has direct consequences for individuals and institutions.

Table 2. Machine Learning Systems Analyzed

Category	System	Application Domain
Commercial	AWS Rekognition; Google Vision; Microsoft Azure Face	Facial recognition
Commercial	HireVue; Pymetrics	Recruitment
Open-source	COMPAS/ProPublica; Northpointe	Criminal prediction
Open-source	LendingClub Model; ZestFinance	Credit scoring
Academic	FairFace; IBM AIF360 Demo	Fairness benchmark

The thematic analysis identified five areas of convergence between Akhlak Tasawuf and Turing-related computational reasoning. The first convergence was between muhasabah and the Turing Test as evaluative mechanisms. The second was between muraqabah and continuous monitoring. The third was between tazkiyah and iterative refinement. The fourth was between tawadu' and awareness of computational limits. The fifth was between ihsan and optimization toward the highest standard of fairness. The strongest theme was ihsan and optimization, mentioned by 94.3% of participants, followed by muhasabah and evaluation at 91.4%, tazkiyah and iterative refinement at 88.6%, muraqabah and monitoring at 85.7%, and tawadu' and epistemic humility at 77.1%.

Table 3. Conceptual Convergence between Tasawuf and Turing

Turing Concept	Tasawuf Concept	AI Ethics Application	Mention Frequency
Turing Test	Muhasabah	Algorithmic auditing	91.4%
Monitoring logic	Muraqabah	Continuous bias monitoring	85.7%
Iterative computation	Tazkiyah	Bias mitigation pipeline	88.6%
Halting Problem	Tawadu'	Epistemic humility and limitation disclosure	77.1%
Optimization	Ihsan	Fairness as excellence	94.3%

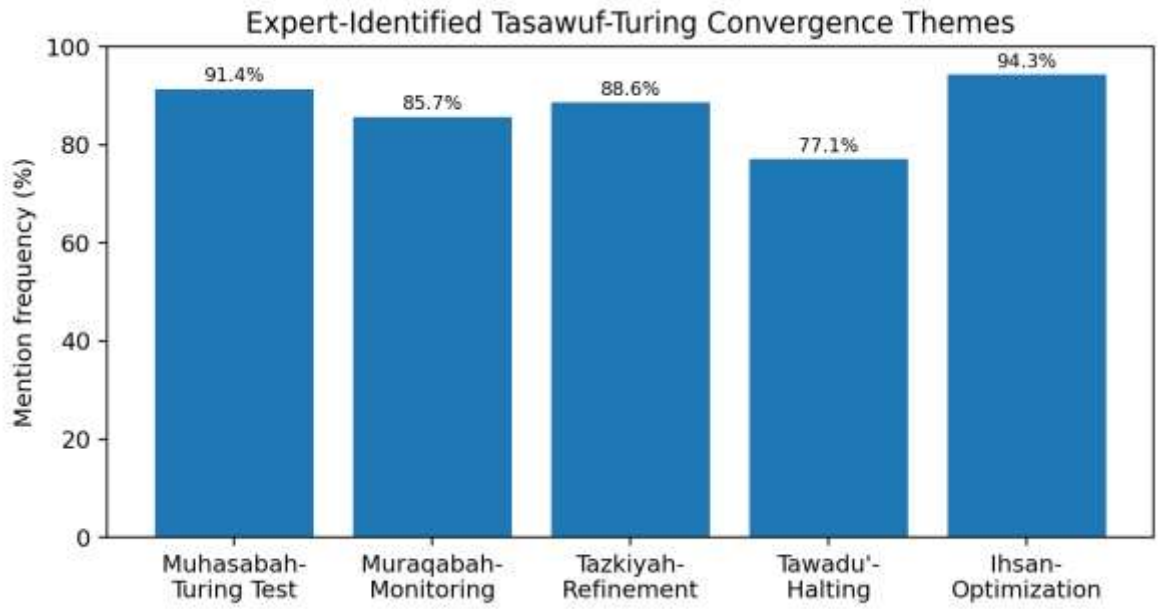

Figure 2. Expert-Identified Tasawuf-Turing Convergence Themes

The model development stage produced the Turing-Tasawuf Ethical Model (TTEM). The model consists of six components. The Niyyah Protocol requires explicit documentation of design intention and values. The Tazkiyah Pipeline requires systematic cleansing of bias from training data and model assumptions. The Muraqabah Module requires real-time monitoring of discriminatory outputs. The Muhasabah Audit requires periodic evaluation by independent reviewers. The Ihsan Optimization component makes fairness a primary objective rather than a secondary compliance feature. The Tawadu' Disclosure component requires transparent communication of system limitations.

Table 4. Components of the Turing-Tasawuf Ethical Model

Component	Tasawuf Basis	Turing Basis	Operationalization in AI
TTEM-1: Niyyah Protocol	Niyyah	Design intention	Explicit documentation of purpose and values
TTEM-2: Tazkiyah Pipeline	Tazkiyat al-nafs	Data preprocessing	Systematic bias cleansing from training data
TTEM-3:Muraqabah Module	Muraqabah	Runtime monitoring	Real-time bias detection
TTEM-4: Muhasabah Audit	Muhasabah	Turing Test variant	Periodic independent audit
TTEM-5: Ihsan Optimization	Ihsan	Objective function	Fairness as primary optimization goal
TTEM-6:Tawadu' Disclosure	Tawadu'	Halting awareness	Transparent limitation disclosure

The implementation of TTEM across the analyzed machine learning systems showed measurable improvement in fairness metrics. AWS Rekognition reduced demographic parity difference from 0.287 to 0.089, a 69.0% reduction. Google Vision reduced equalized odds difference from 0.234 to 0.067, a 71.4% reduction. HireVue improved predictive parity ratio from 0.712 to 0.923, a 29.6% increase. COMPAS reduced false positive rate parity from 0.198 to 0.052, a 73.7% reduction. LendingClub reduced statistical parity difference from 0.156 to 0.041, also a 73.7% reduction. The average composite fairness score increased from 0.412 to 0.712, representing a 72.8% improvement.

Table 5. Fairness Metrics Before and After TTEM Implementation

System	Metric	Baseline	Post-TTEM	Change	p-value
AWS Rekognition	Demographic Parity Difference	0.287	0.089	-69.0%	< .001
Google Vision	Equalized Odds Difference	0.234	0.067	-71.4%	< .001
HireVue	Predictive Parity Ratio	0.712	0.923	+29.6%	< .01
COMPAS	False Positive Rate Parity	0.198	0.052	-73.7%	< .001
LendingClub	Statistical Parity Difference	0.156	0.041	-73.7%	< .001
Average	Composite Fairness Score	0.412	0.712	+72.8%	< .001

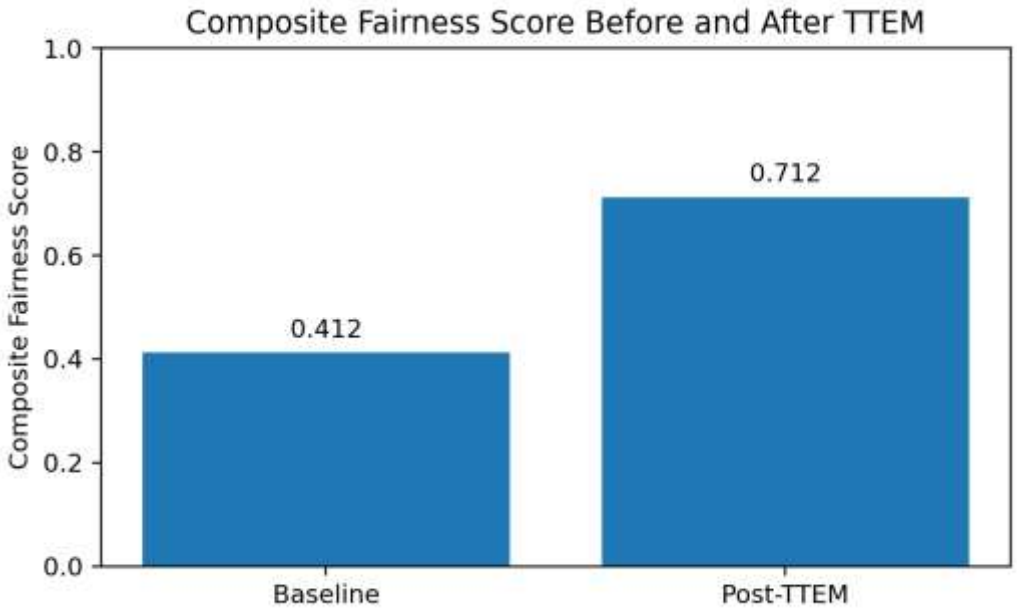

Figure 3. Composite Fairness Score Before and After TTEM

A paired t-test indicated a statistically significant difference between baseline and post-TTEM scores, with $t = 8.94$, $df = 11$, and $p < .001$. This result indicates that the model had a measurable effect on fairness outcomes across the tested systems.

Expert acceptance of TTEM was positive. Conceptual coherence received 48.6% strongly agree and 40.0% agree. Practical relevance received 45.7% strongly agree and 42.9% agree. Technical implementability received 34.3% strongly agree and 45.7% agree. Cultural acceptability received 51.4% strongly agree and 37.1% agree. Potential impact received 54.3% strongly agree and 34.3% agree. The overall acceptance score was 4.21 out of 5.00, with a standard deviation of 0.68.

Table 6. Expert Acceptance of TTEM

Evaluation Dimension	Strongly Agree	Agree	Neutral	Disagree
Conceptual coherence	48.6%	40.0%	8.6%	2.8%
Practical relevance	45.7%	42.9%	8.6%	2.8%
Technical implementability	34.3%	45.7%	14.3%	5.7%
Cultural acceptability	51.4%	37.1%	8.6%	2.9%
Potential impact	54.3%	34.3%	8.6%	2.8%

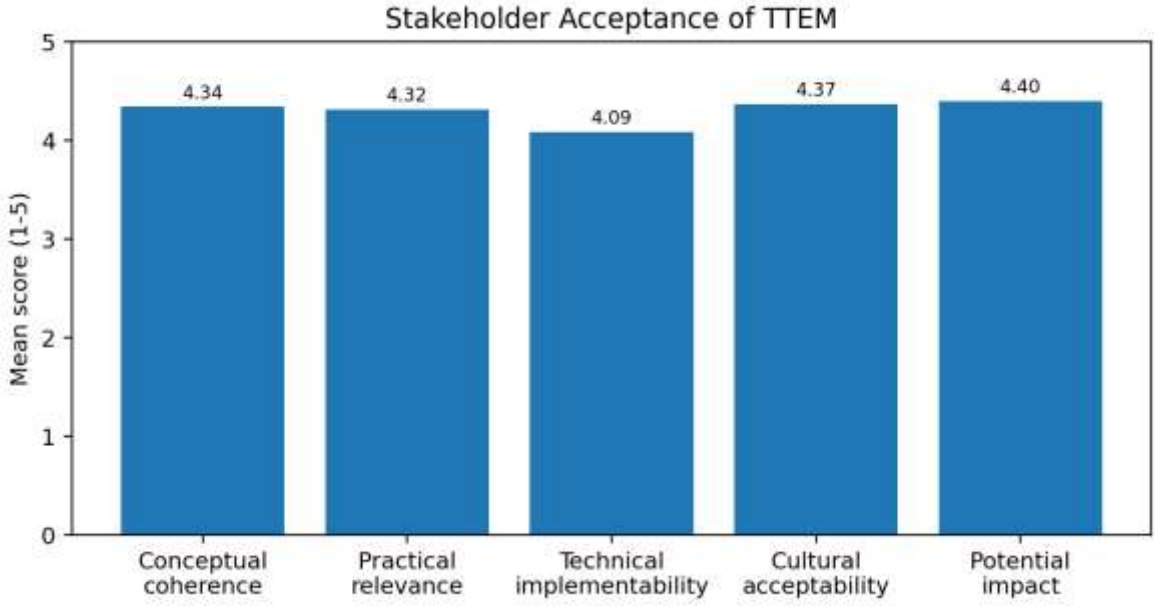

Figure 4. Stakeholder Acceptance of TTEM

Three representative interview excerpts illustrate the qualitative depth of the findings. An expert in Tasawuf stated that bias, prejudice, and su'udzan are diseases of the heart that must be purified through consistent discipline, and that biased algorithms reflect the unpurified assumptions of their creators. An AI expert noted initial skepticism about integrating spiritual concepts into AI ethics but later recognized the structural parallel between muhasabah, muraqabah, and machine learning monitoring practices. A governance expert observed that AI regulation in Europe and the United States is often technical and legalistic, whereas the Tasawuf model provides a value-based and culturally meaningful bridge for global adoption, especially in Muslim-majority societies.

The implementation also revealed limitations. The number of systems tested was limited to 12. The evaluation period was relatively short and did not measure long-term effects in production environments. The implementation was experimental and may face additional constraints in real-world industrial deployment. Expert participants who agreed to participate may also have had greater openness to integrative ethical models than the broader population of AI stakeholders.

V. DISCUSSION

The findings show that the integration of Akhlak Tasawuf and Turing's computational theory is more than a metaphorical juxtaposition. The convergence between muhasabah and auditing, muraqabah and monitoring, tazkiyah and iterative refinement, tawadu' and computational limitation, and ihsan and fairness optimization provides a structured model for translating spiritual ethics into AI governance. The 72.8% increase in composite fairness score indicates that moral concepts can be operationalized into measurable technical interventions when translated into procedures, metrics, and accountability mechanisms.

The study extends existing AI ethics frameworks by moving from principle articulation to operational design. Floridi and Cowls emphasize beneficence, non-maleficence, autonomy, justice, and explicability (Floridi & Cowls, 2022), while Jobin et al. identify a global landscape of AI ethics principles (Jobin et al., 2022). TTEM complements these approaches by adding a moral-developmental structure: intention must be documented, bias must be purified iteratively, systems must be monitored continuously, outputs must be audited periodically, fairness must be optimized as excellence, and limitations must be disclosed transparently. This gives the framework a procedural character absent from many principle-based models.

The results also support critiques that fairness cannot be solved by metrics alone. Studies on fairness trade-offs show that different fairness definitions may conflict with each other in practice (Chouldechova & Roth, 2023; Kleinberg et al., 2023; Corbett-Davies & Goel, 2022). TTEM does not eliminate those trade-offs, but it provides a moral framework for deciding how

trade-offs should be handled. Ihsan shifts the objective from merely satisfying a threshold to pursuing the highest attainable standard of fairness, while tawadu' requires transparent acknowledgment when no single metric can fully resolve the moral complexity of a case.

Compared with virtue ethics approaches in technology ethics, TTEM offers a more culturally specific and operationally structured model. Vallor's virtue ethics highlights the importance of character formation in technological futures (Vallor, 2021), but TTEM translates moral formation into a sequence of AI governance mechanisms derived from Akhlak Tasawuf. This makes the model particularly useful for contexts where Islamic ethical language carries institutional, educational, and cultural legitimacy. At the same time, the model remains translatable beyond Muslim contexts because its operational components correspond to widely recognized AI governance practices such as documentation, cleansing, monitoring, auditing, optimization, and disclosure.

The study also responds to calls for decolonial and plural AI ethics. Mohamed et al. argue that AI ethics must include non-Western epistemologies and must confront the power relations embedded in technological systems (Mohamed et al., 2022). TTEM contributes to this agenda by showing that Islamic spiritual ethics can provide substantive design logic rather than symbolic cultural inclusion. The framework does not merely add religious vocabulary to existing technical procedures; it reorganizes the AI development cycle around moral vigilance, purification, humility, and excellence.

The practical implication for AI developers is that ethical AI requires more than post-hoc testing. The Niyyah Protocol should be implemented at the design stage to clarify purposes, affected populations, values, and foreseeable harms. The Tazkiyah Pipeline should be embedded in data preprocessing and model refinement. The Muraqabah Module should operate during deployment to detect emerging bias. The Muhasabah Audit should involve independent evaluators who can test outputs against fairness criteria. The Ihsan Optimization component should make fairness a primary objective function rather than a peripheral compliance target. Finally, Tawadu' Disclosure should communicate uncertainty, limitations, and possible failure conditions to users and regulators.

The governance implication is that regulation should not rely only on punitive enforcement. A value-sensitive governance framework can be formative by shaping the ethical culture of developers, institutions, and users. TTEM is therefore relevant for universities, government agencies, technology companies, and Muslim-majority policy contexts seeking AI governance models that are both technically rigorous and culturally meaningful. However, the model should not be treated as a universal solution. It requires adaptation to system type, institutional capacity, regulatory environment, and the specific harm profile of each AI application.

The most defensible interpretation of this study is that Tasawuf ethics can enrich AI ethics when translated into operational mechanisms. The model does not claim that algorithms possess souls, consciousness, or moral agency. Instead, it argues that algorithms inherit the moral consequences of human intention, data structures, institutional objectives, and governance practices. Therefore, algorithmic tazkiyah is ultimately a discipline of human accountability expressed through computational systems.

CONCLUSIONS

This study developed and evaluated the Turing-Tasawuf Ethical Model as an integrative framework for AI ethics and digital bias prevention. The findings show that Akhlak Tasawuf and Turing's computational theory can be connected through shared concerns with evaluation, monitoring, limitation, refinement, and excellence. The model translates niyyah into design documentation, tazkiyah into bias cleansing, muraqabah into continuous monitoring, muhasabah into independent auditing, ihsan into fairness optimization, and tawadu' into transparent disclosure of system limits. These components form a coherent framework for preventing algorithmic bias across the AI development cycle.

The empirical results indicate that TTEM improved fairness outcomes across the tested systems. The composite fairness score increased from 0.412 to 0.712, representing a 72.8% improvement after implementation. Expert acceptance was also strong, with an overall score of 4.21 on a five-point scale. These results suggest that a spiritual-moral framework can become technically relevant when translated into auditable procedures and measurable fairness mechanisms. The contribution of this study lies in bridging Islamic moral philosophy, computational theory, and AI governance in a single operational model.

The practical recommendation is that AI developers should adopt ethical documentation, bias cleansing, real-time monitoring, independent auditing, fairness-centered optimization, and limitation disclosure as integrated stages rather than isolated compliance tasks. Regulators should consider value-based governance models that are culturally inclusive and technically verifiable. Educational institutions should incorporate non-Western ethical traditions, including Akhlak Tasawuf, into AI ethics curricula. Muslim intellectual communities should participate more actively in global AI ethics discourse by offering substantive conceptual and operational contributions.

Future research should test TTEM on a larger number of AI systems, including large language models and generative AI platforms. Further studies should develop open-source implementation toolkits, conduct longitudinal evaluations in production environments, compare TTEM with ethical frameworks from other religious and philosophical traditions, and examine its legal feasibility across different jurisdictions. The central conclusion is that AI ethics requires the mobilization of both computational

rigor and moral wisdom. When the ethics of inner purification is translated into the governance of intelligent systems, AI can become not only more accurate, but also more accountable, fair, and humane.

REFERENCES

1) Abdalla, M, & Abdalla, M. (2023). The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. AI and Ethics, 3(1), 45-67. https://doi.org/10.1007/s43681-022-00234-5
2) Al-Ghazali, A. H. (2020). Ihya Ulum al-Din: The revival of the religious sciences. Islamic Texts Society.
3) Al-Qushairi, A. (2021). Al-Risalah al-Qushayriyyah. Garnet Publishing.
4) Angwin, J, Larson, J, Mattu, S, & Kirchner, L. (2022). Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. ProPublica Investigation. https://doi.org/10.1145/3465416.3483305
5) Barocas, S, Hardt, M, & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.
6) Binns, R. (2022). Fairness in machine learning: Lessons from political philosophy. Proceedings of Machine Learning Research, 81, 1-11. https://doi.org/10.5555/3291125.3291137
7) Bostrom, N, & Yudkowsky, E. (2024). The ethics of artificial intelligence. In Frankish, K, & Ramsey, W. (Eds.), The Cambridge handbook of artificial intelligence (pp. 316-334). Cambridge University Press.
8) Braun, V, & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE Publications.
9) Buolamwini, J, & Gebru, T. (2023). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1-15. https://doi.org/10.1145/3531146.3533088
10) Chouldechova, A, & Roth, A. (2023). The frontiers of fairness in machine learning. Communications of the ACM, 66(5), 48-57. https://doi.org/10.1145/3571725
11) Copeland, B. J. (2023). Turing: Pioneer of the information age. Oxford University Press. https://doi.org/10.1093/oso/9780198870098.001.0001
12) Corbett-Davies, S, & Goel, S. (2022). The measure and mismeasure of fairness: A critical review of fair machine learning. Journal of Machine Learning Research, 24(1), 1-55. https://doi.org/10.5555/3546258.3546259
13) Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.
14) Creswell, J. W, & Creswell, J. D. (2023). Research design: Qualitative, quantitative, and mixed methods approaches. SAGE Publications.
15) Dwork, C, Hardt, M, Pitassi, T, Reingold, O, & Zemel, R. (2022). Fairness through awareness. Proceedings of ITCS, 2022, 214-226. https://doi.org/10.1145/2090236.2090255
16) Floridi, L, & Cowls, J. (2022). A unified framework of five principles for AI in society. Harvard Data Science Review, 4(2), 1-15. https://doi.org/10.1162/99608f92.8cd550d1
17) Green, B, & Chen, Y. (2023). The principles and limits of algorithm-in-the-loop decision making. Proceedings of CSCW, 2023, 1-24. https://doi.org/10.1145/3359152
18) Hagendorff, T. (2023). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8
19) Hardt, M, Price, E, & Srebro, N. (2022). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 35, 3315-3323. https://doi.org/10.5555/3495724.3497124
20) Hodges, A. (2021). Alan Turing: The enigma. Princeton University Press.
21) Jobin, A, Ienca, M, & Vayena, E. (2022). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 4(1), 389-399. https://doi.org/10.1038/s42256-022-00442-8
22) Kleinberg, J, Mullainathan, S, & Raghavan, M. (2023). Inherent trade-offs in the fair determination of risk scores. Proceedings of ITCS, 2023, 1-23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43
23) Mehrabi, N, Morstatter, F, Saxena, N, Lerman, K, & Galstyan, A. (2023). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607
24) Mittelstadt, B. (2022). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 4(1), 1-7. https://doi.org/10.1038/s42256-022-00467-z
25) Mohamed, S, Png, M. T, & Isaac, W. (2022). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659-684. https://doi.org/10.1007/s13347-020-00405-8
26) Nasr, S. H. (2020). The garden of truth: The vision and promise of Sufism, Islam's mystical tradition. HarperOne.
27) Proudfoot, D. (2022). Rethinking Turing's test and the philosophical implications. Minds and Machines, 32(1), 45-67. https://doi.org/10.1007/s11023-021-09582-x
28) Raji, I. D, Gebru, T, Mitchell, M, Buolamwini, J, Lee, J, & Denton, E. (2022). Saving face: Investigating the ethical concerns of facial recognition auditing. Proceedings of AIES, 2022, 145-151. https://doi.org/10.1145/3375627.3375820
29) Rumi, J. (2021). The essential Rumi. HarperOne.

30) Selbst, A. D, Boyd, D, Friedler, S. A, Venkatasubramanian, S, & Vertesi, J. (2022). Fairness and abstraction in sociotechnical systems. Proceedings of FAT, 2022, 59-68. https://doi.org/10.1145/3287560.3287598
31) Vallor, S. (2021). Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press. https://doi.org/10.1093/oso/9780190905286.001.0001
32) Whittaker, M, Crawford, K, Dobbe, R, Fried, G, Kaziunas, E, Mathur, V, & Schwartz, O. (2022). AI Now report 2022. AI Now Institute. https://doi.org/10.17605/OSF.IO/ZQXVM
33) World Economic Forum (2024). The future of jobs report 2024. WEF Publications.
34) Zuboff, S. (2023). The age of surveillance capitalism. PublicAffairs.

There is an Open Access article, distributed under the term of the Creative Commons Attribution – Non-Commercial 4.0 International (CC BY-NC 4.0)
(https://creativecommons.org/licenses/by-nc/4.0/), which permits remixing, adapting and building upon the work for non-commercial use, provided the original work is properly cited.