Upholding Amanah: Qur'anic Concept of Trust to Combat AI-Generated Misinformation and Deepfakes

Sania Rahma Harfia
University of Muhammadiyah Surakarta
Email: saniarahmaharfia@gmail.com

Abstract

The development of generative artificial intelligence has increased the production of synthetic content, misinformation, and deepfakes that can disrupt the authenticity of information and trust in digital communication. This study aims to analyze the concept of Qur'anic mandate and reconstruct it as an ethical framework to deal with AI-generated misinformation and deepfakes. The research uses a qualitative approach through an integrative literature review and conceptual analysis of the literature on Islamic information ethics, digital communication, misinformation, deepfakes, digital literacy, and AI governance. Thematic analysis was carried out by grouping findings based on amanah, siddiq, tabayyun, tabligh, technology risk, literacy, and governance mechanisms. The results of the study show that amanah can be placed as the foundation of responsibility in information management, while siddiq regulates the truth of representation, tabayyun regulates information verification, and tabligh regulates responsible delivery. The integration of these four principles results in a framework that can be applied to users, media, educational institutions, digital platforms, and AI developers. The research concludes that the handling of AI-generated misinformation and deepfakes requires integration between Qur'anic ethics, AI literacy, technical verification mechanisms, transparency, and accountability. The trust framework makes a conceptual contribution to the development of Islamic communication ethics and information governance in the era of artificial intelligence.

Keywords: trust, qur'an, artificial intelligence, misinformation, deepfake, tabayyun, information ethics.

Introduction

The development of generative artificial intelligence has changed the way information is produced, replicated, and disseminated in the digital space. Generative technology is no longer limited to the production of text, but is capable of producing synthetic images, sounds, and videos with a level of resemblance that is increasingly difficult to distinguish from authentic content. These developments bring benefits to digital communication, but at the same time increase the risk of misinformation production, information manipulation, and deepfakes. (Al-Kfairy et al., 2024) show that the development of generative artificial intelligence presents ethical issues related to transparency, accountability, and the use of information that can mislead users. (Al-Khazraji et al., 2023) specifically points out that deepfake technology has

implications for the spread of misinformation in social media because it allows the creation of audiovisual representations that resemble real individuals or events. The issue is also related to the social, ethical, economic, and political consequences of misinformation as discussed by (Domenico et al., 2023), while (Singh & Name, 2023) places the increase in deepfake content and AI-based misinformation as a problem that has a global dimension. (Karanja & Nzau, 2026) found that the emergence of deepfakes also affects crisis communication practices because organizations must respond to manipulative content that can undermine public trust. In the context of journalism, the development of AI is also related to information security and the credibility of the investigative process, thus demanding a stronger verification system (Ademola & Somorin, 2024). (Germani et al., 2024) explains the dual nature of AI in the dissemination of information because the same technology can accelerate the distribution of information while also being used to generate or amplify misleading information. (Romanishyn et al., 2025) also places AI-based disinformation as a challenge that requires institutional resilience, policies, and the capacity of society to distinguish between authentic and manipulative information. These conditions show that the problem of AI-generated misinformation and deepfakes is not only technical, but also related to human ethics, trust, and responsibility in producing and distributing information.

The issue of trust is becoming increasingly important because the digital ecosystem depends on the ability of individuals to assess whether the information, sources, and audiovisual representations they receive are trustworthy. (Neyazi et al., 2023) show that information-seeking behavior and concerns about misinformation are related to the level of trust in AI, so the issue of trust cannot be separated from the way people interact with information technology. (Suryani et al., 2023) shows that Islamic values can play a role in building media trust through an emphasis on honesty, responsibility, and the prevention of hoaxes and disinformation. In the Islamic tradition, one of the concepts that has a direct relationship with the issue of trust is trust. (Herijanto, 2022) explained that al-amanah in the Qur'an has a wider scope of meaning than the meaning of trust in modern use because it includes moral responsibility, integrity, and obligation to something entrusted. Daud et al. (2018) placed the trust along with siddiq, fathanah, and tabligh as important characters in Islamic leadership, while Rusyidi (2023) linked these values to governance that is oriented towards integrity and accountability. Chanifah et al. (2024) also placed integrity based on Qur'anic values as an important element in the formation of responsible leadership. The concept of trust does not stand alone because Islamic communication ethics also places siddiq as a demand for honesty, tabayyun as the obligation to verify information, and tabligh as the responsibility to convey information correctly. Bakar et al. (2024) show that tabayyun demands information checking before the information is disseminated, while Riza (2021) connects tabayyun with information literacy skills. Setiawan (2021) shows that the practice of tabayyun also has a social function in preventing conflicts that arise due to the receipt of unverified information. Therefore, the value of trust can be understood as a moral foundation that integrates honesty, verification, responsibility, and information delivery in one framework of Islamic communication ethics.

The literature on Islamic ethics has discussed the application of Qur'anic values in various fields, but the direct relationship between trust and contemporary challenges in the form of AI-generated misinformation and deepfakes still requires more focused conceptual development. (Zubair et al., 2019) has linked the Islamic ethical tradition with efforts to deal with fake news, misinformation, and fake content generated by machine learning, thus

providing a starting foundation for the discussion of Islamic ethics in an AI-based information environment. (Masum et al., 2015) discusses the ethics of information technology based on an Islamic perspective, while (Kausar et al., 2024) discusses the compatibility of Islamic legal principles with the challenges of artificial intelligence. (Fauzi and Ayub, 2019) show how information fiqh can be a guideline for the use of social media through an emphasis on user responsibility and information checking. (Najmuddin and Tajibu, 2023) places efforts to avoid defamation as an important part of Islamic communication ethics, while (Saragih, 2021) emphasizes the role of Islamic journalists in dealing with destructive communication such as hate speech. Islamic value-based communication practices are also developing in digital da'wah journalism (Estuningtyas et al., 2023), the use of new media for da'wah activities (Adeni et al., 2022), and the development of media markets from an Islamic perspective (Amallah et al., 2023). Fadhillah (2022) expands the discussion of Islamic communication ethics in advertising practices, while *Ethics and Morality in Journalism: Islamic Review, an Analysis* (2022) emphasizes the importance of morality in the production and dissemination of journalistic information. (Haron et al., 2020) show that Islamic ethical values have normative characters that can be used in the formation of governance, while Elmanzlawi (2021) shows that Qur'anic values can be recontextualized in contemporary issues outside the field of ritual worship. The study provides a broad conceptual basis, but does not fully integrate mandate as the main framework that specifically links the responsibility of information producers, user verification, technological transparency, and public protection against AI-based deepfakes.

The urgency of this research also arises from the need to connect Islamic normative values with the concrete mechanisms of digital literacy and artificial intelligence governance. Windiarti et al. (2020) place digital literacy in the perspective of Islamic norms as a necessary competency to deal with technological transformation, while Arda and Başarır (2024) highlight the importance of AI literacy in the relationship between humans and artificial intelligence systems. Qur'anic value education has been discussed through character formation and value systems by Mubasyaroh (2013), the development of Islamic education epistemology by Iman (2018), religious character education by Nursikin (2022), and the strengthening of Qur'anic character in the face of global change by Sunardi and Munfarida (2024). Hajjaj and Aimah (2020), Hasanah and Badria (2023), and Ridwan et al. (2024) also show the use of Qur'anic values to form leadership, morals, and behavioral ethos in a changing social context. Samadi et al. (2022) show that Islamic educational institutions are also directly dealing with technological changes in the era of the industrial revolution, while Firdaus (2023) shows the role of the student community in building moderate communication behavior on social media. Nasuki and Askar (2023) even link strategies to dealing with hoaxes with learning Islamic religious education. In the organizational realm, Ibrahim and Tijjani (2024), Setiawan (2023), Suwandi et al. (2019), Maulana et al. (2024), and Abidah et al. (2022) show that the value of trust, Islamic ethics, and governance principles can be applied to organizational responsibility, economic behavior, and relations between stakeholders. Prabangso et al. (2024) and Umar and Qadaruddin (2021) show that the principle of tabayyun and responsible communication also have relevance in solving social problems and delivering religious messages. Based on these gaps, this study aims to reconstruct the concept of Qur'anic trust as an ethical framework to deal with AI-generated misinformation and deepfakes by integrating amanah, siddiq, tabayyun, tabligh, digital literacy, and AI governance. The formulation of this research problem is: how the concept of Qur'anic trust can be constructed as the basis for information ethics in the AI era; how trust, siddiq,

tabayyun, and tabligh can be used to respond to the risk of AI-generated misinformation and deepfakes; and how the framework can be applied to digital literacy, information production, and AI governance in order to strengthen the integrity of digital communication.

Methods

This research uses a qualitative design through an integrative literature review approach and conceptual analysis. The approach was chosen because the purpose of the research is not to measure the statistical relationship between variables, but to develop a theoretical construction regarding the application of trust in dealing with AI-based misinformation and *deepfakes*. The analysis is directed at bringing together Islamic ethical literature with contemporary studies on artificial intelligence, misinformation, digital communication, and technology governance.

The source of research data is in the form of a corpus of literature that discusses amanah, siddiq, tabayyun, tabligh, Islamic communication ethics, misinformation, *deepfakes*, digital literacy, AI literacy, and artificial intelligence governance. Literature is treated as a unit of analysis based on its relevance to the focus of the research. Sources that have identical bibliographic information are treated as one work to avoid double counting.

The analysis instrument is in the form of a literature coding sheet that contains the identity of the source, the focus of the research, the Qur'anic ethical concepts discussed, the form of information threat, prevention mechanisms, responsible parties, forms of literacy, governance mechanisms, and ethical implications. The data collection process is carried out through systematic reading of each source to identify statements related to amanah, siddiq, tabayyun, and tabligh as well as their relationship with misinformation, *deepfakes*, AI technology, digital communication, and public interest.

The literature is then grouped into three broad categories. First, literature on the foundations of Islamic information ethics. Second, literature on the risks of AI-based misinformation and *deepfakes*. Third, literature on the application of Islamic ethics in AI literacy and governance. The grouping follows a conceptual pattern that places trust as the basis of responsibility, siddiq as a demand for truth, tabayyun as a verification mechanism, and tabligh as the principle of responsible information delivery (Bakar et al., 2024; Daud et al., 2018; Herijanto, 2022; Rusyidi, 2023).

Data were analyzed using thematic analysis through the identification of similarities, differences, and conceptual relationships between studies. The synthesis stage was carried out by compiling the normative relationship between individual responsibility, verification processes, information production, and technological governance. The literature on Islamic ethics was compared with research on AI and *deepfakes* to identify the intersection between normative principles and technical needs such as *detection*, *provenance*, *labeling*, transparency, and accountability (Al-Kfairy et al., 2024; Al-Khazraji et al., 2023; Germani et al., 2024).

Results

Amanah as the Foundation of Information Responsibility

The results of the analysis show that trust is a central concept related to responsibility, integrity, reliability, and the obligation to maintain something entrusted. Herijanto (2022) shows that the concept *of al-amanah* in the Qur'an has a broader normative dimension than the

concept of trust in the general sense. Trust does not only describe the relationship of trust, but requires individuals to account for something that is within their authority.

In the context of digital information, something that is entrusted does not have to be a physical object. Data, information, access to technology, the ability to produce content, and the authority to manage digital platforms can also be understood as a form of responsibility. Users who have access to AI systems have responsibility for how the technology is used. Developers have responsibility for the system designed. Platforms have responsibility for the distribution and moderation mechanisms of information. The recipient of information also has responsibility for the decision to forward or stop the dissemination of information.

These findings are in line with the study of trust in organization and governance. Daud et al. (2018) linked trust to the quality of Islamic leadership, while Rusyidi (2023) linked it to good governance. Chanifah et al. (2024) placed Qur'anic integrity as an important element in leadership. Although the research does not directly address *deepfakes*, the principles of responsibility and integrity put forward can be used as a normative foundation for digital information governance.

Thus, trust in an AI environment can be reconstructed as a moral obligation to maintain the integrity of information during the process of creation, processing, storage, distribution, and use. Responsibility does not stop with content creators, but is spread throughout the digital communication chain.

Siddiq, Tabayyun, and Tabligh as Operational Principles

Amanah requires operational principles so that it can be applied in communication practice. The results of the analysis show that siddiq, tabayyun, and tabligh have complementary functions.

Siddiq sets standards of truth and honesty. In the context of AI, this principle demands that technology not be used to form intentionally false representations to mislead others. Suryani et al. (2023) show that honesty and accountability are important elements in building trust in the media. This principle is particularly relevant to *deepfakes* because manipulative content can create the impression as if someone is saying or doing something that actually never happened.

Tabayyun functions as an epistemic mechanism. Bakar et al. (2024) place tabayyun as a process of checking information before dissemination. Riza (2021) connects this concept with information literacy. In the digital context, tabayyun means that users do not automatically receive information just because the visual or audio form looks convincing. Users need to check the source, context, consistency of information, and supporting evidence.

Tabayyun in an AI environment can also be translated into technical mechanisms. Checks can be done through metadata tracing, content origin checks, original source searches, the use of *deepfake detection tools*, watermark identification, and *cross-checking* with credible sources. Thus, verification technology does not contradict the concept of tabayyun, but it can be a contemporary instrument for carrying out the principle.

Tabligh regulates the dimensions of delivery. Correct information must still be delivered responsibly. The delivery of information needs to consider the context, impact, public interest, and potential losses that may arise. Estuningtyas et al. (2023) show that Islamic communication practices continue to evolve in the digital environment so that its ethical demands need to keep up with changes in the media.

These four principles form a systematic relationship. Trust becomes a moral orientation. Siddiq establishes the integrity of the content. Tabayyun regulates the process of examination. Tabligh regulates the process of delivery. This combination makes Qur'anic ethics not stop at abstract norms, but can be translated into concrete stages in information management.

The Risks of AI-Based Misinformation and Deepfakes

The analysis shows that AI-based misinformation has a different character than conventional misinformation. AI allows manipulative content to be produced quickly, cheaply, in large quantities, and with a high level of realism. Al-Khazraji et al. (2023) show that *deepfakes* pose challenges in terms of detection, misinformation, and social implications.

Al-Kfairy et al. (2024) also identified various ethical challenges to the use of generative AI and emphasized the importance of multidisciplinary solutions. These problems cannot be solved through technological accuracy alone. When a system is able to detect manipulative content, there are still questions about who is responsible for the production of that content, who should make corrections, and how the resulting losses should be handled.

Germani et al. (2024) affirm the dual nature of AI in the dissemination of information. AI can accelerate access to knowledge, but at the same time facilitate the production of false information. Romanishyn et al. (2025) point out that the response to AI-based disinformation requires institutional policies and capacity. Karanja and Nzau (2026) also point out that the emergence of *deepfakes* can affect crisis communication because organizations must respond to manipulative material that can undermine public trust.

The findings suggest that responding to AI-based misinformation requires a combination of *detection*, *verification*, *provenance*, transparency, literacy, and governance. Detection systems are important, but they cannot replace human responsibility. It is in this context that trust provides a normative dimension that detection technology cannot provide.

Literacy and Governance as a Form of Implementation

The next group of findings is related to the application of Qur'anic values to education and technology governance. Windiarti et al. (2020) show the relevance of Islamic norms in the development of digital literacy. Arda and Başarır (2024) place AI literacy as an important competency for a society that is increasingly interacting with artificial intelligence.

Trust-based AI literacy does not only teach how to recognize synthetic content. It also develops an awareness of responsibility when receiving, producing, and disseminating information. Users need to understand that technology's ability to produce content does not automatically make all uses of technology ethically acceptable.

Nasuki and Askar (2023) show that strategies to deal with hoaxes can be integrated into the learning process of Islamic Religious Education. These findings open up space for the development of tabayyun-based AI literacy modules. Education can teach students to recognize synthetic media, examine sources, *cross-check*, understand the limitations of detection tools, and restrain the dissemination of information when the truth cannot be ascertained.

At the institutional level, trust can be translated into transparency, accountability, traceability, and protection of the public interest. Zubair et al. (2019) show that the Islamic ethical tradition can provide insight in dealing with *fake news*, misinformation, and fake content generated by *machine learning*. Kausar et al. (2024) also emphasized the need to read the development of AI through the perspective of Islamic law and ethics.

Discussion

The results of the study show that amanah has the potential to be placed as the center of the information ethics framework in the era of artificial intelligence. This concept goes beyond the notion of trust in interpersonal relationships. Amanah emphasizes the obligation to account for something received, managed, produced, or delivered. This meaning is consistent with Herijanto (2022), who shows that *al-amanah* has a broader normative scope than the concept of *trust*.

This meaning is important in the context of *deepfakes*. The main issue does not only lie in whether or not people believe a video. Ethical questions also include who creates the content, why the content is created, who distributes it, whether the platform provides a marker for synthetic content, whether the recipient conducts checks, and who is responsible when the content results in harm.

The trust framework allows these responsibilities to be distributed to every actor in the information ecosystem. Developers have a mandate to design and implement systems with regard to potential abuse. Users have a mandate not to use AI to produce misleading manipulations. The media has a mandate to conduct verification. Platforms have a mandate to build a system of transparency, reporting, correction, and mitigation. Recipients of information have a mandate not to spread unchecked content.

An important contribution of this framework lies in the transformation of the mandate from individual value to the principle that organizes the relationship between people, technology, organization, and the public interest. The mandate is not placed as a substitute for regulation or technology. Rather, it provides a normative basis for the use of such regulations and technologies.

The siddiq principle clarifies the dimension of information integrity. In the context of generative AI, truth does not only mean true factual statements. Truth is also related to representation. An image or video may appear realistic, but it is still misleading if it depicts an event that never happened. Therefore, the use of AI to create synthetic representations should be accompanied by adequate explanations if there is a possibility that the public will consider it authentic documentation.

Suryani et al. (2023) show that honesty and accountability are important elements in the formation of media trust. In this context, siddiq can be translated as a prohibition on making false claims regarding content and technological capabilities. The principle also supports the obligation of developers to explain the limitations of AI systems, including possible errors, *hallucinations*, or manipulations.

Tabayyun provides a more concrete mechanism. Bakar et al. (2024) and Riza (2021) place tabayyun as a verification process that is closely related to information literacy. In the past, tabayyun could be done through confirmation to the source or examination of testimonials. In the digital environment, the same principle can be applied through the examination of primary sources, metadata, authentication marks, *reverse search*, provenance, watermarks, and detection tools.

This expansion does not change the substance of tabayyun. Instead, technology has become a new instrument to carry out the same epistemic demands, namely not making decisions based on unclear information. Setiawan (2021) shows that tabayyun has a social function in preventing conflicts that arise due to the receipt of information without examination.

This function is even more important when *deepfakes* can visualize fake events in a form that seems very convincing.

Tabligh complements this process with a dimension of communication responsibility. Not all information that is technically correct needs to be disseminated without considering the context. Information needs to be conveyed accurately, proportionately, and considering the consequences. On social media, the act of sharing content is not a neutral action because each distribution can increase the reach of the message. Therefore, users are not only responsible for the content they create themselves, but also for the content they amplify.

The integration of amanah, siddiq, tabayyun, and tabligh results in a sequential framework. Before information is produced, the amanah provides orientation regarding goals and responsibilities. At the production stage, siddiq demands the integrity of representation. When information is received, tabayyun demands verification. When information is to be forwarded, tabligh demands responsible delivery. This framework can be applied to both individual and institutional users.

The significance of such frameworks becomes increasingly apparent when compared to the technological approach to *deepfakes*. Al-Khazraji et al. (2023) point out the importance of detection, while Al-Kfairy et al. (2024) point out the need for multidisciplinary solutions to generative AI challenges. Detection systems can generate estimates of whether content is manipulated. However, they cannot normatively determine whether someone should have created fake videos to deceive others.

Thus, responses that rely solely on detection technology are inadequate. Conversely, responses that rely solely on moral prohibitions without the support of verification mechanisms are also ineffective. Germani et al. (2024) show that AI can be part of the problem as well as part of the solution. Based on these findings, this study places trust as a normative foundation that works alongside technical mechanisms.

At the literacy level, these implications result in a need for an AI literacy model that integrates technical and ethical competencies. The ability to recognize the signs of *deepfakes* is important, but users also need to have an ethical orientation to stop the spread of information when its authenticity cannot be ascertained.

Arda and Başarır (2024) emphasize the importance of AI literacy in the human relationship with artificial intelligence systems. Meanwhile, Windiarti et al. (2020) point out the possibility of integrating Islamic norms in digital literacy. The two approaches can be brought together through a trust-based literacy model.

Literacy at least includes an understanding of the character of synthetic media, the ability to examine sources, cross-verification skills, an understanding of the limitations of detection systems, awareness of algorithmic bias, and responsibilities before disseminating content. Through this approach, users not only ask, "Is this content genuine?", but also, "Do I have sufficient grounds to believe it and can I account for the decision to distribute it?"

At the governance level, trust can be translated into a number of practical principles. First, transparency regarding the use of AI. Users should be informed when certain content is generated or modified using AI if the information is relevant to the meaning of the content.

Second, the mechanism *of provenance* or traceability of the origin of the content. Information about the source, modification history, and production process can help the public assess authenticity. Third, correction and reporting mechanisms. Platforms need to allow users to report manipulative content and obtain explanations of actions taken.

Fourth, accountability. Institutions need to determine who is responsible if AI systems generate or expand the spread of harmful information. Fifth, clear *labeling* of synthetic content. Labels do not always solve the problem of manipulation, but they can reduce the likelihood of synthetic content being perceived as authentic documentation.

The principle is consistent with the need for transparency and accountability discussed by (Al-Kfairy et al., 2024), Germani et al. (2024), and Romanishyn et al. (2025). In contrast, the trust framework provides a normative basis based on Qur'anic ethics that explains why transparency, verification, and accountability are necessary.

Thus, the main theoretical contribution of this research is to reconstruct the mandate of individual moral values into a multi-layered ethical framework. The first layer is on individuals as users and producers of information. The second layer is on organizations, media, and educational institutions. The third layer is on digital platforms and artificial intelligence developers. Each layer has a different form of responsibility, but all are centered on the protection of information integrity and the public interest.

The framework also shows that Islamic ethics should not be placed in opposition to modern technology. Qur'anic values can be recontextualized as normative principles, while technology provides a mechanism for its implementation. Tabayyun can use detection tools. Trust can be realized through accountability. Siddiq can be applied through *disclosure* and labeling. Tabligh can be applied through transparent communication standards.

Framework for Implementing Mandates in the AI Ecosystem

Based on the synthesis of the research, the implementation of the mandate framework can be formulated in four main dimensions.

First, trust as a responsibility. Each actor is responsible for his or her role in the information cycle. The developer is responsible for the design and risk of the system. The platform is responsible for distribution. The media is responsible for verification. The user is responsible for the production and dissemination.

Second, siddiq as the integrity of representation. Content should not be presented in a way that deliberately creates a false perception of a fact, a person's identity, or an event. Synthetic content needs to be accompanied by a caption if it has the potential to be understood as authentic content.

Third, tabayyun as verification. Users and institutions need to develop an examination process that is in accordance with the level of risk of information. Information that has major public consequences requires stronger verification than low-risk information.

Fourth, tabligh as responsible communication. The delivery of information needs to consider the truth, context, way of presentation, and potential social consequences.

These four dimensions must be supported by digital literacy and governance. Without literacy, it is difficult for users to run tabayyun. Without platform transparency, it is difficult for users to check sources. Without accountability, trust is only an individual demand without institutional support. Therefore, this framework requires the integration of individuals, education, technology, and policies.

Conclusion

This research shows that the concept of Qur'anic trust can be developed into a relevant ethical framework to deal with artificial intelligence-based misinformation and *deepfakes*. Trust is not only understood as trust, but as a moral responsibility to the process of receiving, producing, verifying, using, and disseminating information.

The framework becomes more operational when the mandate is integrated with siddiq, tabayyun, and tabligh. Siddiq establishes demands for the correctness and honesty of information representation. Tabayyun provides the principle of examination before information is received or disseminated. Tabligh emphasizes responsibility when information is conveyed to other parties. These four principles form an ethical system that includes the purpose, content, verification process, and process of conveying information.

The literature on generative AI and *deepfakes* shows that technological developments have increased the scale, speed, and realism of manipulative content so that the authenticity of information can no longer be judged solely through user intuition (Al-Kfairy et al., 2024; Al-Khazraji et al., 2023; Germani et al., 2024). Therefore, the trust framework needs to be integrated with technical mechanisms such as detection, *provenance*, watermarking, labeling, transparency, and verification.

The main contribution of the research lies in the reconstruction of amanah as a layered ethical principle that is not only aimed at individuals. Amanah also includes the media, educational institutions, organizations, digital platforms, and artificial intelligence developers. Each actor has a responsibility to maintain the integrity of information in accordance with its role in the digital ecosystem.

At the individual level, the trust framework encourages users not to generate or disseminate manipulative content. At the education level, the framework can be used in the development of tabayyun-based AI literacy. At the media level, the trust encourages the strengthening of verification procedures. At the platform and developer level, the trust can be realized through transparency, *disclosure*, traceability, correction mechanisms, and accountability.

Further research needs to test the framework empirically. Researchers can develop instruments to measure the understanding of amanah, siddiq, tabayyun, and tabligh and test their relationship with the ability to recognize and contain the spread of *deepfakes*. Experiments can also compare the effectiveness of general AI literacy with amanah and tabayyun-based AI literacy.

Further research can also develop trust-based platform governance indicators that include transparency, *disclosure*, *labeling*, *provenance*, correction, and accountability. In the context of Islamic education, research can develop modules that integrate the ability to detect synthetic media with Qur'anic verification ethics. Cross-cultural and cross-country studies are also needed to understand variations in the implementation of mandates in different social and regulatory environments.

Overall, the protection of information integrity in the age of artificial intelligence requires a combination of human moral responsibility and technical mechanisms. Detection technology without ethics cannot determine how technology should be used, whereas ethics without verification capabilities struggle to respond to the complexity of synthetic media. Amanah, siddiq, tabayyun, and tabligh provide a normative foundation that can connect the two needs. With further development, the mandate framework can move from normative concepts to

digital ethical models that can be measured, tested, and applied in artificial intelligence literacy and governance.

Acknowledgments

I would like to express our sincere gratitude to University for their unwavering support and encouragement throughout this research. Additionally, I extend my heartfelt thanks to the anonymous reviewer for providing invaluable feedback and insightful suggestions that greatly contributed to the improvement of this paper.

Conflict of interests

The author declares no conflicts of interest.

References

Al-Kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics, 11*(3), 58. https://doi.org/10.3390/informatics11030058

Al-Khazraji, S., Saleh, H. H., Khalid, A., & Mishkhal, I. (2023). Impact of deepfake technology on social media: Detection, misinformation and societal implications. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 23*, 429–441. https://doi.org/10.55549/epstem.1371792

Arda, Z., & Başarır, L. (2024). Artificial intelligence literacy and the demand for augmented intelligence. *AdComunica*, 115–142. https://doi.org/10.6035/adcomunica.8020

Bakar, A. Y. A., Fazial, F., Abdullah, S. R., Taher, N. H. M., Kasim, N. H., & Hamid, M. S. A. (2024). The concept of tabayyun in dissemination of information through mass media. *International Journal of Academic Research in Business and Social Sciences, 14*(1). https://doi.org/10.6007/IJARBSS/v14-i1/20693

Chanifah, N., Samsudin, A., & Ansori, I. H. (2024). Quranic leadership: The effort to realize the integrity of leaders in preventing corruption. *International Journal of Educational Management and Innovation, 5*(1), 24–38. https://doi.org/10.12928/ijemi.v5i1.9173

Daud, W. N. W., Rahim, M. A., Zainol, F. A., & Ismail, A. G. (2018). Enhancing Takaful operator performance through the Islamic leader's quality. *Ar-Rā'iq, 1*(1), 1–42. https://doi.org/10.59202/riq.v1i1.440

Domenico, G. D., Tiago, M. T. B., Viglia, G., & Cheng, Y. (2023). Guest editorial: The social, ethical, economic and political implications of misinformation. *Internet Research, 33*(5), 1665–1669. https://doi.org/10.1108/INTR-10-2023-947

Estuningtyas, R. D., Hamid, A., Nurhaidah, S. N., & Majid, Z. A. (2023). The role of the millennial generation in da'wah journalism in the digital age. *Jurnal Syntax Transformation, 4*(7), 35–45. https://doi.org/10.46799/jst.v4i7.758

Fauzi, N. A. F., & Ayub, A. (2019). Fikih Informasi: Muhammadiyah's perspective on guidance in using social media. *Indonesian Journal of Islam and Muslim Societies, 9*(2), 267–293. https://doi.org/10.18326/ijims.v9i2.267-293

Germani, F., Spitale, G., & Biller Andorno, N. (2024). The dual nature of AI in information dissemination: Ethical considerations. *JMIR AI, 3*, e53505. https://doi.org/10.2196/53505

Herijanto, H. (2022). Al amanah in al Qur'an vs trust: A comparative study. *International Journal of Ethics and Systems, 38*(4), 549–575. https://doi.org/10.1108/IJOES-03-2021-0064

Karanja, J. M., & Nzau, T. (2026). The emergence of AI deepfakes and their impact on strategic crisis communication: Practitioners' perspectives from Kenya. *Journal of Contingencies and Crisis Management, 34*(1). https://doi.org/10.1111/1468-5973.70140

Kausar, S., Leghari, A. R., & Soomro, A. S. (2024). Analysis of the Islamic law and its compatibility with artificial intelligence as an emerging challenge of the modern world. *Annals of Human and Social Sciences, 5*(I). https://doi.org/10.35484/ahss.2024(5-i)10

Masum, A. K. M., Ullah, M. C., & Azad, M. A. K. (2015). Information technology (IT) ethics in the light of Islam. *IIUC Studies, 9*, 243–260. https://doi.org/10.3329/iiucs.v9i0.24030

Nasuki, A., & Askar, R. A. B. (2023). Overcoming hoax: Prophet Muhammad's strategies integrated in PAI teaching and learning. *El-Tarbawi, 16*(1), 57–92. https://doi.org/10.20885/tarbawi.vol16.iss1.art3

Neyazi, T. A., Ee, T. K., Nadaf, A. H., & Schroeder, R. (2023). The effect of information seeking behaviour on trust in AI in Asia: The moderating role of misinformation concern. *New Media & Society, 27*(4), 2414–2433. https://doi.org/10.1177/14614448231212804

Riza, M. (2021). Information literacy as an implementation of Tabayun concept in Islam. *Tik Ilmeu: Journal of Library and Information Science, 5*(2), 201. https://doi.org/10.29240/tik.v5i2.2866

Romanishyn, A., Malytska, O., & Goncharuk, V. A. (2025). AI-driven disinformation: Policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence, 8*. https://doi.org/10.3389/frai.2025.1569115

Rusyidi, I. (2023). Good governance according to Islamic perspective. *Munaddhomah: Journal of Islamic Education Management, 4*(4), 1001–1007. https://doi.org/10.31538/munaddhomah.v4i4.727

Setiawan, D. E. (2021). The significance of Tabayyun practice as conflict resolution in Indonesian society. *Islam Realitas: Journal of Islamic & Social Studies, 7*(2), 181. https://doi.org/10.30983/islam_realitas.v7i2.4654

Suryani, W., Kango, A., & Akifah, A. (2023). Constructing trust in media through Islamic values: Countering hoaxes and disinformation. *Al-Ulum, 23*(1), 21–39. https://doi.org/10.30603/au.v23i1.3441

Windiarti, I. S., Norcahyono, N., & Prabowo, A. (2020). Digital literacy for the millennial generation in Industrial Revolution 4.0 era in Islamic norms perspective. https://doi.org/10.4108/eai.27-8-2020.2303180

Zubair, T., Raquib, A., & Qadir, J. (2019). Combating fake news, misinformation, and machine learning generated fakes: Insights from the Islamic ethical tradition. *ICR Journal, 10*(2), 189–212. https://doi.org/10.52282/icr.v10i2.42