Theology Journal
Online ISSN: 2783-0667 Print ISSN: 2476-5325
Homepage: pke.journals.miu.ac.ir

Artificial Intelligence and the Risk of Distortion of Religious Concepts: An Analysis of Data-Driven, Algorithmic, and Textual Roots

Ruhullah Mowahedi[1]✉ and **Mohammad Khavari**[2]

1. *Corresponding Author*, Research Fellow, Imam Reza International Research Center, Al-Mustafa International University, Mashhad, Iran.Email: r.mowahedi@gmail.com
2. Visiting Researcher, Department of Philosophy of Religion, mam Reza International Research Center, Al-Mustafa International University, Mashhad, Iran. Email: khavary200009@gmail.com

Article Info

Article type:
Research Article

Article history:
Received
19 October 2025
Received in revised form
27 November 2025
Accepted
21 December 2025
Available online
22 December 2025

Keywords:
Artificial intelligence,
religious education,
data bias,
algorithmic bias,
complexity of religious texts

ABSTRACT

Although the discussion of AI errors is a frequent topic in computer science, the analysis of its consequences for the distortion of religious concepts has yet to receive serious attention. Aiming to address this gap, the present article offers a theoretical framework for identifying the roots of and managing AI errors in this domain. According to this framework, such errors are the product of the interaction of three interrelated factors: (1) bias in data: the structural marginalization of Shīʿī sources in training data; (2) bias in algorithms: technical deficiencies and value-orientations of model designers; and (3) the complexity of religious texts: the resistance of metaphorical language and multi-source textual structures to mechanical comprehension. In line with this framework, the article emphasizes the necessity of employing a systemic set of technical and strategic solutions, including the creation of benchmark corpora and knowledge graphs at the data layer; the use of fine-tuning, retrieval-augmented generation, and human feedback at the algorithmic layer; and a paradigm shift toward the "illuminative tool" at the textual layer. Finally, by distinguishing functions into three categories—content-based (high-risk), formal (safe), and dual (conditional)—the article demonstrates that stringent technical solutions are reserved for the content-based domain, whereas in other domains, focusing on enhancing digital literacy and skills such as prompt engineering suffices.

Cite this article: Mowahedi, R., & Khavari, M. (2025). Artificial Intelligence and the Risk of Distortion of Religious Concepts: An Analysis of Data-Driven, Algorithmic, and Textual Roots. *Theology Journal, 12*(2), 31-48. https://doi.org/10.22034/pke.2026.22375.2048.

 Publisher: Al-Mustafa International University.
DOI: 10.22034/pke.2026.22375.2048

Introduction

Artificial intelligence (AI), as one of the transformative technologies of the contemporary era, has opened new horizons in numerous fields, including education, thereby raising the possibility of its integration into religious education. However, transferring this technology into a sensitive and multi-layered domain such as religion does not merely present opportunities; it also brings forth a host of epistemological challenges. Among these, the most concerning issue is AI's capacity to generate distorted responses regarding religious teachings. Language models may provide rulings based on unreliable sources when answering jurisprudential questions; rely on weak or marginal narrations in recounting historical events; and present superficial or incompatible interpretations with Islamic and Shi'a foundations when explaining theological concepts.

Since diagnosis precedes treatment, analyzing the roots of these errors is of fundamental importance. Despite this significance, systematic research in this area—especially from an interdisciplinary perspective that bridges the technical and religious dimensions of the issue—is still in its infancy. In the Persian literature, existing studies are either limited to general education or, adopting a descriptive and optimistic view, neglect a critical analysis of the challenges. In English literature, although the volume of research is larger, the focus remains predominantly on general education; in the rare instances where religious education is addressed, the issue of response errors has not been seriously investigated from the perspective of Islamic, particularly Shi'a, traditions. Aiming to bridge this profound research gap, the present study seeks to provide a theoretical framework for the structural root-cause analysis of errors, while outlining a path for transitioning from naive optimism to a responsible approach in utilizing AI.

Methods

Rather than conducting case-by-case examinations of errors, this study analyzes them based on the three core components of processing systems. Since every text processing system possesses three inherent components—input (data), processing (algorithm), and the object of processing (the nature of the text)—any internal error is inevitably traced back to a deficiency or complexity in one of these three components, or their interaction. This paper endeavors to investigate each of these three error factors in relation to religious teachings.

Findings

The research findings indicate that the primary structural root of errors is data bias, or systemic input deficiency. Feeding AI models with quantitatively and qualitatively unbalanced data contaminates the model's epistemological foundation from the outset. This bias operates across three dense layers: In the

first layer, due to the dominance of secular and capital-driven discourses on the web, religious concepts as a whole are marginalized by massive amounts of entertainment and technical data. In the second layer, because technology development hubs are concentrated in the West and translation movements in the Islamic world remain weak, Islamic concepts are underrepresented compared to Western and Christian ones. In the third layer, due to the minority status of the Shi'a population and the heavy investments of Sunni institutions in digitizing their heritage, the volume of digital resources available for model training vastly exceeds that of Shi'a sources. This quantitative disparity causes AI to learn the Sunni narrative as the standard representation of Islam, thereby neglecting or distorting specific Shi'a concepts.

The second factor, which operates independently of data, is algorithmic bias or processing errors. Even if the data were corrected, the inherent logic of language models—designed around statistical probabilities to generate fluent text rather than to discover truth—remains a source of error. This logic causes these models, on the one hand, to engage in uncritical acceptance, reproducing any linguistically coherent data without content validation. On the other hand, due to this generation-oriented logic, they tend toward "hallucination" in the absence of information, generating fabricated yet plausible answers instead of remaining silent. Ultimately, algorithms carry the latent values of their designers and may promote liberal concepts, such as absolute freedom of speech, in ways that clash with religious boundaries and the sanctity of sacred figures.

The third root factor is the inherent complexity of religious texts, representing a challenge in the object of processing. The language of religion is multi-layered, metaphorical, and symbolic; AI, lacking lived experience and human intuition, often falls into literal and superficial interpretations when encountering it, failing to grasp its semantic depth. Moreover, religious knowledge possesses a multi-source structure, the correct comprehension of which requires managing conflicts among narrations and formulating ijtihadi (jurisprudential reasoning) syntheses—a complex skill that currently lies beyond the linear and statistical processing capabilities of AI.

Conclusion

Concluding the threefold analysis, the paper demonstrates that securing AI in the religious domain cannot be achieved through one-dimensional solutions; rather, it requires a constellation of combined measures across all three layers. In the data layer, it is essential to move beyond mere digitization toward constructing "Gold Standard Corpora" and "Shi'a Knowledge Graphs." In the algorithmic layer, technical challenges must be addressed using Retrieval-Augmented Generation (RAG) and Reinforcement Learning from Human Feedback (RLHF), while mitigating developers' value biases demands moving toward Supervised Fine-Tuning (SFT), Domain-Adaptive Training

(DAT), and the development of indigenous models. In the text layer, the paradigm of human-machine interaction must shift from answering to illumination (using "Illuminator Tools").

The paper's primary strategic contribution is the risk-based classification of AI functions: content-based functions are identified as high-risk areas requiring strict supervision, whereas formal functions are presented as a safe domain for utilization, and dual functions remain conditional upon user expertise. This intelligent classification prevents technology's opportunities from being overlooked under the pretext of risks, while ensuring that the authenticity of religious knowledge is not sacrificed to oversimplification.

Declarations

Authors' Contributions: Both authors contributed equally to the conception, design, and drafting of this work.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable corrective feedback.

Ethical Considerations: The authors have observed all ethical principles in conducting and publishing this scientific research, and this has been confirmed by all of them.

Funding: This research was funded by the Imam Reza International Research Center, Al-Mustafa International University (Khorasan Branch).

Conflict of Interest: The authors declare no conflict of interest.

Generative AI and AI-Assisted Technologies Statement: In the preparation of this paper, artificial intelligence was utilized solely for language editing and improving readability; all analyses and conclusions remain the sole responsibility of the authors.