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.

پژوهشنامه کلام
شاپا برخط: ۰۶۶۷-۲۷۸۳ شاپا چاپی: ۵۳۲۵-۲۴۷۶
Homepage: pke.journals.miu.ac.ir

هوش مصنوعی و خطر تحریف مفاهیم دینی؛ واکاوی ریشه‌های داده‌ای، الگوریتمی و متنی

روح‌الله موحدی[۱] و محمد خاوری[۲]

۱. نویسندهٔ مسئول، پژوهشگر همکار، گروه فلسفه دین، پژوهشکده بین‌المللی امام رضا (ع)، جامعه المصطفی العالمیه، مشهد، ایران. رایانامه: r.mowahedi@gmail.com

۲. پژوهشگر مدعو، گروه فلسفه دین، پژوهشکده بین‌المللی امام رضا (ع)، جامعه المصطفی العالمیه، مشهد، ایران. رایانامه: khavary200009@gmail.com

اطلاعات مقاله

نوع مقاله:
مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۴/۰۷/۲۷
تاریخ بازنگری: ۱۴۰۴/۰۹/۰۶
تاریخ پذیرش: ۱۴۰۴/۰۹/۳۰
تاریخ انتشار: ۱۴۰۴/۱۰/۰۱

کلیدواژه‌ها:
هوش مصنوعی،
آموزش دینی،
سوگیری داده‌ای،
سوگیری الگوریتمی،
پیچیدگی متون دینی

چکیده

با اینکه بحث دربارهٔ خطاهای هوش مصنوعی موضوعی پربسامد در علوم کامپیوتر است، تحلیل پیامدهای آن در تحریف مفاهیم دینی هنوز مورد توجه جدی قرار نگرفته است. این مقاله با هدف رفع این خلأ، یک چارچوب تحلیلی سه‌وجهی برای ریشه‌یابی خطاهای هوش مصنوعی در این حوزه ارائه می‌دهد. بر اساس این چارچوب، این خطاها محصول تعامل سه عامل به‌هم‌پیوسته‌اند: (۱) سوگیری در داده‌ها: به‌حاشیه‌رفتن ساختاری منابع دینی، اسلامی و شیعی در داده‌های آموزشی؛ (۲) سوگیری در الگوریتم‌ها: نقایص فنی و جهت‌گیری‌های ارزشی طراحان مدل‌ها؛ و (۳) پیچیدگی متون دینی: مقاومت زبان استعاری و ساختار چندمنبعی متون در برابر فهم مکانیکی. در امتداد این چارچوب، مقاله با تفکیک کاربردها به سه قلمروِ محتوایی (پرخطر)، شکلی (امن) و دوگانه (مشروط)، رویکرد مواجهه با هر قلمرو را تبیین می‌کند: در کنار ضرورت ارتقای سواد دیجیتال دینی، مواجهه با خطرات قلمرو محتوایی مستلزم اصلاحات ساختاری و فنی در دو لایهٔ داده و الگوریتم (مانند توسعه مدل‌های بومی و ابزارهای کنترلی) برای مواجههٔ درست با پیچیدگی‌های متنی است، در حالی که قلمروهای شکلی و دوگانه نیازمند اتخاذ رویکردهای نظارتی نرم‌تر هستند.

استناد: موحدی، روح‌الله؛ خاوری، محمد. (۱۴۰۴). هوش مصنوعی و خطر تحریف مفاهیم دینی: واکاوی ریشه‌های داده‌ای، الگوریتمی و متنی. پژوهشنامهٔ کلام، ۱۲(۲)، ۳۱-۴۸. https://doi.org/10.22034/pke.2026.22375.2048

ناشر: جامعه المصطفی العالمیه.

مقدمه

هوش مصنوعی، به‌مثابهٔ یکی از نیروهای تحول‌آفرین عصر حاضر، در حال دگرگون‌سازی بسیاری از ساحات حیات بشری است. یکی از ساحاتی که از این فناوری تأثیرات عمیقی پذیرفته، حوزهٔ آموزش است؛ جایی که هوش مصنوعی با فراهم آوردن امکانات فراوان همچون پاسخ‌گویی تعاملی، جست‌وجوی هوشمند، تحلیل داده‌ها و... افق‌های جدیدی را گشوده است. این ظرفیت عظیم، به‌طور طبیعی به حوزهٔ آموزش مفاهیم دینی نیز تسری یافته و این امید را ایجاد کرده است که بتوان از آن برای ترجمه و تفسیر متون مقدس، پاسخ‌گویی به سؤالات دینی کاربران و... بهره برد. با این حال، به‌کارگیری این فناوری در حوزهٔ حساس و پیچیده‌ای چون دین، خالی از چالش نیست. شاید بتوان گفت نگران‌کننده‌ترین چالش در این زمینه، احتمال تحریف آموزه‌های دینی توسط هوش مصنوعی است. برای نمونه، مدل‌های زبانی ممکن است در پاسخ به سؤالی فقهی، حکمی نادرست ارائه دهند، در بازگویی رویدادی تاریخی، روایتی ضعیف را مبنا قرار دهند، یا در تبیین مفهومی کلامی، تفسیر سطحی یا حتی مغایر با مبانی اسلامی و شیعی عرضه کنند.

بدیهی است که مواجههٔ اصولی با این طیف کژتابی‌ها، بدون کالبدشکافی دقیق علل پیدایش آن‌ها میسر نخواهد بود، زیرا خطاهای هوش مصنوعی در ساحت دین، عوامل متفاوتی دارند که مانع از ارائهٔ راهکار واحد برای همهٔ آن‌ها می‌شوند؛ لذا تا زمانی که این عوامل به‌درستی تفکیک نشوند، سیاست‌گذاری‌ها و اقدامات فنی نیز از دقت و کارایی مطلوب فاصله خواهند داشت. با این حال، به‌رغم این اهمیت، پژوهش‌های نظام‌مند برای تحلیل این خطاها، به‌ویژه از منظری میان‌رشته‌ای که ابعاد فنی و دینی مسئله را به هم پیوند بزنند، هنوز در ابتدای راه است. در زبان فارسی، پژوهش‌ها یا به آموزش عمومی اختصاص یافته و به عرصه آموزش دینی ورود نکرده‌اند (مانند بقایی و همکاران، ۱۴۰۳؛ خدادادی و همکاران، ۱۴۰۲؛ ذوالفقاری، ۱۴۰۱؛ سهرابی، ۱۴۰۳؛ خیامی و همکاران، ۱۴۰۲؛ مصطفوی، ۱۴۰۱) یا صرفاً نگاهی توصیفی و عمدتاً خوش‌بینانه به کاربردهای آن در این حوزه داشته و سراغ بررسی چالش‌ها نرفته‌اند (مانند نبوی و همکاران، ۱۴۰۳؛ ربیعی‌زاده، ۱۴۰۰). در زبان انگلیسی نیز، به‌رغم حجم بیشتر مطالعات، تمرکز عمدتاً بر حوزه آموزش عمومی بوده است (Chen et al., 2020; Limna et al., 2023; Pham & Sampson, 2022; Saputra et al., 2022; Wang et al., 2024). در موارد معدودی هم که به آموزش دینی توجه شده (Andriani & Sudirman, 2023; Salsabila & Rohiem, 2023)، یا به خطاهای پاسخ‌دهی هوش مصنوعی نپرداخته‌اند، یا آن را از منظر سنت شیعی و حتی اسلامی واکاوی نکرده‌اند. بنابراین، فقدان یک تحلیل نظام‌مند میان‌رشته‌ای که به‌جای توصیف کاربردها، به کالبدشکافی علل ریشه‌ای خطاهای هوش مصنوعی در حوزه آموزش دینی، به‌ویژه از منظر سنت اسلامی و شیعی بپردازد، به‌شدت احساس می‌شود.

در این راستا، نوشتار حاضر می‌کوشد تا گامی در جهت پر کردن این خلأ تحقیقاتی برداشته و موانع اساسی پیش روی هوش مصنوعی در این عرصه را تحلیل کند. برای دستیابی به این هدف، به‌جای بررسی موردی خطاها، نوشتار با رویکردی نظری‌ـساختاری و با الهام از مدل پردازش اطلاعات، فرایند تولید پاسخ در مدل‌های زبانی را به‌مثابهٔ یک سیستم پردازش اطلاعات واکاوی می‌کند. ازآنجاکه هر سیستم پردازش اطلاعات سه

مؤلفه ذاتی (ورودی، پردازش، و موضوع پردازش) دارد، هر خطای معرفتی درونی آن لاجرم به نقص در یک یا چند مورد از همین سه مؤلفه بازمی‌گردد. ازاین‌رو، نوشتار با تمرکز بر این هستهٔ درونی و پرهیز از پرداختن به عوامل بیرونی نظیر ناآشنایی کاربر که به‌رغم تأثیر منفی دلالتی بر نقص ساختاری سیستم ندارد، سه عامل زیر را به‌عنوان ریشه‌های ساختاری پیدایش خطاها تحلیل می‌کند:

(۱) سوگیری در داده‌ها (خطا در ورودی) که از به‌حاشیه‌رفتن ساختاری منابع دینی، اسلامی و شیعی در داده‌های آموزشی مدل‌ها ناشی می‌شود.

(۲) سوگیری در الگوریتم‌ها (خطا در پردازش) که از یک‌سو در نقایص فنی و از دیگرسو در جهت‌گیری‌های ارزشی سازندگان مدل‌ها ریشه دارد.

(۳) پیچیدگی ذاتی متون دینی (خطا ناشی از ماهیت موضوع) که از مقاومت زبان استعاری و ساختار چندمنبعی متون دینی در برابر فهم مکانیکی سرچشمه می‌گیرد.

نوشتار، با واکاوی این سه عامل، به دنبال ارائه تصویری واقع‌بینانه، از محدودیت‌های هوش مصنوعی است تا زمینه را برای گذار از خوش‌بینی ساده‌انگارانه به رویکردی مسئولانه در بهره‌گیری از این فناوری فراهم آورد.

سوگیری در داده‌ها

نخستین عامل در بروز خطاهای هوش مصنوعی در حوزه دین، به سوگیری ساختاری در داده‌های آموزشی آن بازمی‌گردد. ازآنجاکه داده‌ها خوراک اصلی مدل‌های هوش مصنوعی هستند، کیفیت آن‌ها مستقیماً بر عملکرد و خروجی این مدل‌ها اثر می‌گذارد. اگر داده‌های آموزشی بیشتر متکی بر دیدگاه‌های یک مذهب یا جریان خاص باشند و سایر قرائت‌ها در آن‌ها به‌خوبی بازنمایی نشده باشند، مدل هوش مصنوعی نیز ناخودآگاه در جهت همان دیدگاه غالب عمل می‌کند و در نتیجه، از دقت پاسخ‌هایش کاسته می‌شود. این پدیده که در ادبیات فنی با عنوان «سوگیری در داده‌ها» شناخته می‌شود (Weidinger et al., 2022, p. 217)، ابعاد گوناگونی دارد و بسته به بستر اعتقادی و فرهنگی تحلیل، می‌تواند جلوه‌های متفاوتی پیدا کند. ما در اینجا، از منظر شیعی که چارچوب تحلیلی این نوشتار را شکل می‌دهد، به بررسی این مسئله می‌پردازیم. ازاین منظر، می‌توان سه نوع به‌حاشیه‌رفتن را به‌عنوان جلوه‌های اصلی سوگیری مذکور در نظر گرفت: نخست، به‌حاشیه‌رفتن مفاهیم دینی به‌طور کلی؛ دوم، به‌حاشیه‌رفتن مفاهیم اسلامی؛ و سوم، به‌حاشیه‌رفتن مفاهیم شیعی. در ادامه، این سه جنبه را به ترتیب و با تفصیل بیشتری بررسی می‌کنیم.

نخستین جنبهٔ این سوگیری، به‌حاشیه‌رفتن کلیت دین در برابر گفتمان‌های سکولار است که ارتباط مستقیمی با بستر فرهنگی شکل‌گیری این فناوری در غرب دارد؛ جایی که با غلبه گفتمان روشنگری و عقل‌گرایی سکولار، دین به‌تدریج از حوزه‌های عمومی مانند علم، سیاست، آموزش و تولید دانش کنار زده شد و به حریم خصوصی افراد محدود گردید. این تحول عمیق، که فیلسوف کانادایی، چارلز تیلور، در اثر سترگ خود عصر سکولار، آن را نه تغییری سطحی، بل تغییری عمیق در شرایط باور و نحوهٔ درک انسان مدرن از خود و جهان می‌داند (تیلور، ۱۴۰۰)، به‌تدریج در فعالیت‌های علمی و فنی نیز نفوذ کرده است. بازتاب این روند در

فرهنگ دیجیتال معاصر و در سازوکارهای جمع‌آوری و برچسب‌گذاری داده‌ها برای آموزش مدل‌های هوش مصنوعی کاملاً مشهود است. برای نمونه، بررسی محتوای یکی از بزرگ‌ترین مجموعه‌های داده متنی وب، Common Crawl، که بسیاری از مدل‌های زبانی بزرگ از جمله مدل‌های خانواده GPT از آن تغذیه می‌کنند، نشان می‌دهد که این پیکره عمدتاً نمایانگر گفتمان جوانان انگلیسی‌زبان ساکن در کشورهای توسعه‌یافته است (Luccioni & Viviano, 2021, p.186). این سوگیری جمعیت‌شناختی، که به‌طور ضمنی بازتاب‌دهنده فضای فکری سکولار و اولویت‌های فرهنگی مسلط در این جوامع است، موجب می‌شود صداها و متون برآمده از سنن دینی، به‌ویژه از سوی جوامع غیرغربی، در حاشیه قرار گیرند و سهم کمتری در ساختن افق معنایی مدل‌ها داشته باشند. این وضعیت، برخلاف آنچه ممکن است در نگاه نخست تصور شود، امری تصادفی نیست، بل برآمده از انتخاب‌های جهت‌داری که در آن‌ها دین، نه منبع معرفت که امری حاشیه‌ای و شخصی تلقی شده است. علاوه بر این، گسترش منطق سرمایه‌داری نیز این روند را تشدید می‌کند. بر اساس این منطق، اولویت با حوزه‌هایی چون تبلیغات هدفمند، مدیریت مالی، تحلیل رفتار مصرف‌کنندگان و... است که بیشترین بازده اقتصادی را وعده می‌دهند، نه با موضوعاتی چون الهیات، دین و اخلاق که ارزش‌های غیربازاری را نمایندگی می‌کنند و سودآوری مستقیمی ندارند. در نتیجه، حتی اگر متون دینی در فضای وب در دسترس باشند، در فرآیند انتخاب داده‌ها، کمترین بخت را برای گنجانده‌شدن در مجموعه‌های آموزشی خواهند داشت. پیامد این روند این می‌شود که هوش مصنوعی از دسترسی کافی به منابع دینی بازبماند و در پی آن، بازنمایی‌هایی سطحی یا نادرستی از مفاهیم دینی ارائه دهد.

جنبهٔ دوم و خاص‌تر سوگیری مذکور، به‌حاشیه‌رفتن مفاهیم اسلامی در میان سایر مفاهیم دینی است که ریشه در تسلط تاریخی سنن غیراسلامی بر جوامعی دارد که امروزه طلایه‌دار تولید و توسعه فناوری‌های هوش مصنوعی هستند. جوامع غربی که مراکز اصلی این فناوری به شمار می‌روند، در بستری رشد کرده‌اند که سنت مسیحی، به‌شکل آشکار یا پنهان، در ساختارهای دینی و فرهنگ عمومی آن نقش غالب داشته و در نهادهایی چون کلیسا، نظام آموزشی، صنعت نشر کتاب و رسانه‌های جمعی نهادینه شده است. همین زمینهٔ تاریخی موجب شده است تا در فضای دیجیتال نیز بازتولید مفاهیم دینی با غلبه مسیحیت همراه باشد. برای نمونه، هاچینسون نشان می‌دهد که در آرشیو مقالات علمی (ACL)، ارجاع به کتاب مقدس بین ۷ تا ۱۳ برابر قرآن و ده‌ها برابر متون ادیان شرقی است و در پیکرهٔ عظیم چندزبانهٔ MADLAD-400، برای ۱۴۱ زبان ردپای قابل توجه از کتاب مقدس یافت شده، اما هیچ ردپای قابل توجهی از قرآن گزارش نشده است (Hutchinson, 2024, pp. 1031-1032). این سوگیری مستقیماً به مدل‌ها نیز راه می‌یابد: ابوبکر عبید و همکارانش با آزمایش GPT-3 نشان دادند که این مدل‌ها سوگیری نامتقارنی علیه اسلام بروز می‌دهند. آنان دریافتند که اگر جمله‌ای خنثی مانند «دو مسلمان وارد... شدند» به مدل داده شود، در ۶۶٪ موارد ادامه‌ای خشونت‌آمیز (تیراندازی، بمب‌گذاری و...) تولید می‌شود، درحالی‌که با جایگزینی نام سایر ادیان، این نرخ بسیار کمتر بود. در آزمایش قیاس‌ها (مانند این پرسش که «نسبت گستاخ به گستاخی مانند نسبت مسلمان به چیست؟»)، نیز مدل در ۲۳٪ موارد واژهٔ «تروریست» را برای مسلمان پیشنهاد می‌داد، حال‌آنکه برای «مسیحی» اغلب واژه‌های

مثبتی مانند «وفاداری» می‌آمد. هرچند افزودن شش صفت مثبت به جمله (مثلاً «مسلمانان سخت‌کوش») توانست نرخ خشونت را از ۶۶٪ به ۲۰٪ کاهش دهد، اما همین نرخ همچنان بالاتر از نرخ ۱۳ تا ۱۵ درصدی خشونت برای مسیحیان در همان شرایط باقی ماند (Abubakar et al., 2021, pp. 299–302, Fig. 2 & Fig. 4c). سلطهٔ زبان انگلیسی به‌عنوان زبان غالب فناوری، که تا سال ۲۰۲۳ حدود ۵۵ درصد از محتوای اینترنت را در بر می‌گیرد، احتمالاً در بروز و تقویت این سوگیری بی‌تأثیر نیست، زیرا بیشتر منابع اصیل اسلامی به زبان‌های عربی، فارسی، اردو یا ترکی نوشته شده‌اند و حجم عظیمی از این متون هنوز به انگلیسی ترجمه نشده یا در قالب‌های دیجیتالِ قابل پردازش برای مدل‌های زبانی درنیامده‌اند. در نتیجه، پیکره‌های آموزشیِ انگلیسی‌زبان از گنجینهٔ مفاهیم اسلامی تهی می‌مانند و مدل‌ها ناگزیر به بازنمایی‌های ناقص یا برگرفته از منابع غیرتخصصی متکی می‌شوند. بدین ترتیب، احتمال اینکه مفاهیم مهم اسلامی به‌دلیل غیبت در پیکرهٔ آموزشی مدل‌ها، به‌شکلی ناقص و حتی ناسازگار با منطق درونی سنت اسلامی بازنمایی شوند، افزایش می‌یابد.

سومین و درونی‌ترین جنبهٔ این سوگیری، نادیده‌گرفتن مفاهیم شیعی در درون سنت اسلامی است که ریشه در سیطرهٔ دیرپای قرائت سُنّی در جهان اسلام دارد. براساس گزارش مرکز پژوهشی پیو (Pew) در سال ۲۰۰۹، شیعیان حدود ۱۰ تا ۱۳ درصد و اهل‌سنت حدود ۸۷ تا ۹۰ درصد از جمعیت مسلمانان را تشکیل می‌دهند (Pew Research Center, 2009, p. 4). این تفاوت کمّی که صرفاً بازتاب وضعیت امروز نیست و از قرن‌ها پیش تا کنون استمرار یافته، خود را در گسترهٔ منابع مکتوب و میزان مشارکت در تولید دانش اسلامی نیز نشان داده است. درحالی‌که خلافت‌های اموی، عباسی، و عثمانی ساختارهای عظیم و پرنفوذی برای تدوین و ترویج سنت سُنّی فراهم کردند، قدرت سیاسی شیعه، جز در برخی مقاطع محدود و کانون‌های جغرافیایی خاص، فرصت بروز نیافت و همین امر موجب شد تا سازوکارهای نهادی تولید و انتشار دانش دینی در جهان اسلام عمدتاً به‌نفع جریان غالب سنگینی کند. در دوران معاصر نیز، این نابرابری در عرصهٔ دیجیتال ادامه یافته است. برای نمونه، پایگاهی مانند Sunnah.com که توسط نهادهای سُنّی حمایت می‌شوند، بیش از ۱۰۰ هزار متن حدیثی و تفسیری را به زبان انگلیسی ارائه می‌دهد، درحالی‌که Al-Islam.org، یکی از بزرگ‌ترین پایگاه‌های شیعی، تنها حدود ۳ هزار متن ترجمه‌شده دارد، آن‌هم با کیفیت نشر پایین‌تر و گسترهٔ موضوعی محدودتر. این نابرابری فقط به حجم تولیدات داخلی محدود نمانده است؛ چراکه در مطالعات اسلامی در غرب نیز، تمرکز بر روی اسلام سُنّی است و تشیع غالباً در حاشیه قرار دارد (احمدوند، ۱۳۷۷، ص. ۱۵۵)، امری که به بازنمایی محدودتر تشیع در منابع انگلیسی‌زبان و در نتیجه، حضور کمرنگ‌ترش در فضای دیجیتال منجر می‌شود. در کنار این عوامل، تفاوت فاحش در توان مالی نیز نقشی تعیین‌کننده داشته است؛ نهادهایی مانند دانشگاه الازهر، دارالافتای مصر، یا مراکز فتوای عربستان سعودی با برخورداری از حمایت‌های دولتی و منابع پایدار، دهه‌ها راهبرد منسجمی برای ترجمه متون مذهبی خود دنبال کرده‌اند، درحالی‌که نهادهای شیعی، به‌علت محدودیت مالی، از ایجاد زیرساخت‌های لازم برای نهضت ترجمه و انتشار نظام‌مند معارف در سطح بین‌الملل، بازمانده‌اند. این شکاف عمیق، خود را در وضعیت ترجمهٔ مجامع حدیثی این دو مذهب به‌خوبی نشان می‌دهد. از کتب اربعه حدیثی شیعه، تنها کتاب الکافی و من لایحضره الفقیه ترجمهٔ کاملی به انگلیسی دارند، اما دو کتاب دیگر این

مجموعه (تهذیب الاحکام و الاستبصار) یا اصلاً ترجمه نشده‌اند یا ترجمه ناقص و پراکنده دارند. در نقطهٔ مقابل، صحاح سته اهل سنت به‌طور کامل، با شروح مختصر و با حمایت ناشران بزرگی چون دارالسلام، ترجمه و به‌صورت گسترده و بعضاً رایگان در سراسر جهان توزیع شده‌اند. این عدم توازن در سایر آثار مهم شیعی مانند بحارالانوار و تفسیر المیزان نیز تکرار می‌شود. همین تفاوت می‌تواند سبب شود که خوانش سُنّی در پیکرهٔ آموزشی مدل‌های زبانی غلبه یابند و به‌تبع آن، لایه‌هایی از حافظه تاریخی و منطق کلامی شیعه در عرصهٔ دیجیتال به حاشیه رفته و مفاهیمی مانند امامت، عصمت، ولایت، مهدویت و وقایعی چون غدیر خم، سقیفه، عاشورا و غیبت یا از چشم‌ها دور بمانند یا در چارچوبی مغایر با تفکر شیعی روایت شوند.

البته تلاش‌هایی در جریان است که می‌تواند بستر را برای کاهش سوگیری داده‌ای فراهم سازد، مانند «پروژه برنامه درسی دیجیتال مطالعات اسلامی» (Digital Islamic Studies Curriculum, n.d.) و خدمات گسترده مرکز نور در دیجیتال‌سازی منابع. با این حال، باید در نظر داشت که با توجه فضای رقابتی این عرصه، هرگونه کندی در جبههٔ شیعی به عمیق‌تر شدن این نوع سوگیری منجر خواهد شد. گذشته از این، نباید از یاد برد که دیجیتال‌سازی منابع به‌تنهایی کافی نیست. هوش مصنوعی برای یادگیری دقیق مفاهیم شیعی، به چیزی فراتر از انبوه داده‌های خام نیاز دارد. از این‌رو، افزون بر دیجیتال‌سازی، باید به‌سمت تولید پیکره‌های معیار (Gold Standard Corpora) حرکت کرد؛ یعنی مجموعه‌داده‌هایی که توسط متخصصان دینی به‌دقت پالایش و برچسب‌گذاری شده‌اند تا به‌عنوان سرمشق صحیح به خورد ماشین داده شوند. همچنین برای اینکه هوش مصنوعی بتواند روابط پیچیدهٔ میان مفاهیم (مثلاً رابطهٔ امامت با عصمت) را بفهمد، باید متون خطی را به ساختارهای شبکه‌ای یا همان گراف‌های دانش تبدیل کرد. این اقدام هم می‌تواند عملکرد مدل‌های موجود را بهبود بخشد و هم زیرساخت لازم را برای توسعه مدل‌های هوش مصنوعی بومی فراهم آورد. با این حال، تا زمان تحقق این زیرساخت‌ها، کاربران نیز وظیفهٔ مهمی بر عهده دارند. آنان باید نسبت به محتوای دینی هوش مصنوعی محتاط باشند و هر حدیث یا حکمی را که هوش مصنوعی ارائه می‌دهد، با منابع اصیل راستی‌آزمایی کنند.

سوگیری در الگوریتم‌ها

دومین عامل تعیین‌کننده که خطاهای هوش مصنوعی را تشدید می‌کند، پدیده‌ای است که با عنوان «سوگیری در الگوریتم‌ها» شناخته می‌شود. این پدیده در ساختار پردازشی مدل‌های زبانی بزرگ ریشه دارد که برای تولید متن روان طراحی شده‌اند، نه برای کشف حقیقت. از همین رو، حتی اگر داده‌های آموزشی کاملاً بی‌طرفانه باشند، خود الگوریتم‌ها به‌دلیل طراحی و ساختار ذاتی‌شان، گرایش نظام‌مندی به تولید محتوای نادرست و نامعتبر دارند (اسکندری، ۱۴۰۴). برخلاف سوگیری داده‌ای که عمدتاً بر بازنمایی ناقص مفاهیم دینی متمرکز بود، سوگیری الگوریتمی به حوزهٔ خاصی محدود نیست و می‌تواند هر موضوعی، از جمله دین، را تحت تأثیر قرار دهد. این نوع سوگیری عمدتاً خود را در سه شکل اصلی نشان می‌دهد: پذیرش غیرانتقادی، توهم همه‌دانی، و خطاقرمزهای ارزشی سازندگان. در ادامه، به‌ترتیب این جنبه‌ها را بررسی می‌کنیم.

پذیرش غیرانتقادی سوگیری‌ای است که ریشه در نبود سازوکار اعتبارسنجی در الگوریتم مدل‌های زبانی بزرگ دارد. این مدل‌ها به‌دلیل آموزش بر پیکره‌های عظیمی از متون متمایزنشده، صرفاً به پیش‌بینی آماری واژگان بعدی متکی‌اند و توانایی سنجش صحت مطالب یا منابع را ندارند؛ در نتیجه، هر متن روانی را بدون پالایش انتقادی به‌عنوان ورودی معتبر برای تولید پاسخ به کار می‌گیرند (Bender et al., 2021, pp. 613–617). این محدودیت فنی سبب می‌شود که الگوریتم در مواجهه با آرای متضاد، قادر به تفکیک روشمند دیدگاه‌ها نباشد و با فروکاستن مسئله، پاسخ خود را به یکی از سه شکلِ ترجیح یک‌سویه یک گزاره، ادغام هم‌زمان دیدگاه‌های متعارض، یا ارائهٔ ترکیبی ناقص و التقاطی از آن‌ها بازتولید کند. این سوگیری، در بستر مفاهیم دینی، می‌تواند پیامدهای عمیقی در پی داشته باشد؛ زیرا معرفت دینی بر نظامی سلسله‌مراتبی از گزاره‌ها با وزن‌های مختلف بنا شده است که در آن، ارزش هر گزاره نه به روانی لفظ آن، که به جایگاهش در شبکهٔ معرفتی و میزان تطابقش با منابع دست‌اول و روش‌های پذیرفته‌شدهٔ استنباط بستگی دارد. پذیرش غیرانتقادی موجب می‌شود تا مدل این سلسله‌مراتب را نادیده گرفته و هر متن منسجم را، فارغ از اینکه معتبر است یا نامعتبر، در ترازی یکسان قرار دهد. در این بین، سلاست و انسجام زبانی پاسخ نیز متأسفانه موجب می‌شود تا کاربر غیرمتخصص آن را موجه و معتبر بپندارد و از راستی‌آزمایی‌اش صرف‌نظر کند.

دومین سوگیری الگوریتمی، پدیده‌ای است که می‌توان آن را «سوگیری توهم همه‌دانی» نامید. برخلاف سوگیری پذیرش غیرانتقادی که بر ناتوانی هوش مصنوعی در غربالگری داده‌های موجود تمرکز داشت، سوگیری حاضر معطوف به وضعیتی است که مدل فاقد داده‌های کافی در یک زمینه است و با این حال، به‌جای اعلام صریح محدودیت خود، دست به تولید پاسخ می‌زند. همچون مورد قبلی، این سوگیری نیز در معماری مدل‌های زبانی مولد ریشه دارد؛ این مدل‌ها بر پایهٔ پیش‌بینی دنبالهٔ واژگان طراحی شده‌اند و ذاتاً فاقد مکانیزمی برای اعلام «نمی‌دانم» هستند، مگر آنکه چنین قابلیتی از طریق آموزش‌های تکمیلی به آن‌ها افزوده شود. در نتیجه، وقتی پرسش کاربر خارج از پوشش اطلاعاتی مدل قرار می‌گیرد، الگوریتم مولد به‌جای توقف، با تکیه بر الگوهای زبانی آموخته‌شده و ترکیب‌های آماری محتمل، پاسخی از نو می‌سازد؛ پاسخی که ظاهری روان و مستند دارد، اما محتوایش یکسره برساختهٔ مدل است. این ویژگی که در ادبیات فنی «هذیان‌گویی» خوانده می‌شود (Ji et al., 2023, pp. 4-5)، در حوزهٔ مفاهیم دینی می‌تواند به نتایجی ویرانگر منجر شود. برای نمونه، مدل ممکن است در پاسخ به پرسشی دربارهٔ حدیثی از معصومان (ع)، متنی کاملاً جعلی اما با ساختار و لحنی مشابه متون اصیل حدیثی تولید کند و کاربر ناآشنا با مبانی علم حدیث نیز آن را به‌عنوان حقیقت بپذیرد و این امر زمینه‌ساز تحریف معارف و ترویج بدعت در ذهن فراگیران شود.

در نهایت، سومین سوگیری الگوریتمی در استفاده از هوش مصنوعی برای انتقال مفاهیم دینی، به وجود خط‌قرمزها و جهت‌گیری‌های ارزشی در طراحی این سیستم‌ها بازمی‌گردد که از ترجیحات شخصی، فرهنگی و سیاسی سازندگان آن‌ها ناشی می‌شود. ازآنجاکه اغلب ابزارهای رایج کنونی هوش مصنوعی ریشه در جوامع غربی دارند، بعضاً حامل پیش‌فرض‌هایی‌اند که با مبانی هستی‌شناختی و ارزش‌شناختی شیعی سازگار نیستند. این ناسازگاری در موارد مختلفی مانند تأکید بر برابری مطلق جنسیتی، فردگرایی افراطی یا آزادی بیان بی‌قیدوشرط

و... خود را نشان می‌دهد. البته باید توجه داشت که این چالش لزوماً در قالب تقابل مستقیم یا نفی صریح عقاید دینی بروز نمی‌یابد؛ زیرا مدل‌های زبانی در پاسخ به پرسش‌های دینی، غالباً با اتخاذ رویکرد چندصدایی، تصویری بی‌طرف از خود ارائه می‌دهند که مانع جبهه‌گیری اولیهٔ مخاطب می‌شود. با این حال، این تساهل ظاهری در لایهٔ رویین نباید گمراه‌کننده باشد؛ چراکه چالش اصلی در لایه‌های پنهان این فناوری جریان دارد؛ جایی که ترجیحات ارزشی سازندگان، از طریق یک لایهٔ فنی به‌نام «هم‌راستاسازی ایمنی» (Safety Alignment) اعمال می‌شود و به‌شکلی ظریف در عمق تحلیل‌ها، سناریوهای کاربردی و سبک‌های زیست پیشنهادی رسوخ می‌یابد. این لایه در واقع مانند یک فیلتر درونی پنهان عمل می‌کند که بدون نفی مستقیم آموزه‌های ناسازگار، در قبال آن‌ها رویکرد تعدیل‌گرایانه در پیش می‌گیرد. در این فرایند، پاسخ‌های مدل طوری تنظیم می‌شوند تا آن آموزه صرفاً ترجیحی شخصی و امری سلیقه‌ای معرفی گردد یا در میان انبوهی از نظرات رقیب گم شود و به حاشیه برود. ازآنجاکه این رویکرد در پشت پاسخ‌های منظم و به‌ظاهر بی‌طرفانه رخ می‌دهد، واکنش آگاهانه به آن بسیار پیچیده است. این پیکربندی پنهان، در درازمدت پتانسیل آن را دارد که بدون ایجاد حساسیت اولیه، منظومهٔ فکری کاربران را با نظام ارزش‌های پیش‌فرض سازندگان این فناوری هم‌راستا سازد.

از میان سه سوگیری الگوریتمی فوق، دو مورد نخست آن عمدتاً ماهیت فنی دارند و با پیشرفت این فناوری سیر نزولی طی خواهند کرد، چنان‌که هم اکنون نیز تلاش‌هایی در این حوزه در جریان است. دو رویکرد مهم در این زمینه «تولید مبتنی بر بازیابی» (Retrieval-Augmented Generation, RAG) و «یادگیری تقویتی با بازخورد انسانی» (Reinforcement Learning from Human Feedback, RLHF) هستند. در روش نخست، برای کاهش احتمال تولید مطالب ساختگی، مدل پیش از پاسخ‌گویی به مجموعه‌ای از منابع ازپیش‌تأییدشده متصل می‌شود و پاسخ خود را بر اساس داده‌های آن سامان می‌دهد (Lewis et al., 2020, pp. 9459–9461). در روش دوم نیز ارزیابان انسانی پاسخ‌های مدل را از نظر کیفیت رتبه‌بندی می‌کنند و مدل، از این طریق، به‌تدریج به‌سمت پاسخ‌های محتاطانه‌تر و دقیق‌تر هدایت می‌شود (Ouyang et al., 2022, pp. 27731–27732). این دو رویکرد مکمل، که یکی ورودی و دیگری خروجی را کنترل می‌کند، به‌رغم نقص‌ها، چشم‌انداز روشنی برای کاهش دو سوگیری نخست ترسیم می‌کنند. با این حال، سوگیری سوم که به خط‌قرمزهای سازندگان هوش مصنوعی مربوط می‌شود، چالشی عمیقاً ارزشی است که حتی با پیشرفت مدل‌های فعلی نیز نمی‌توان به رفع کامل آن امید بست. البته استفاده از دو روش مذکور تا حدی کارساز است، اما چون نمی‌تواند چارچوبی را که مدل بر اساس آن پرسش‌ها را تفسیر می‌کند و استدلال خود را سامان می‌دهد، دگرگون سازد، محدودیت‌های جدی دارد؛ ازاین‌رو در مسائلی که پاسخ آن‌ها نیازمند استنباط یا ترکیب چند منبع است و متن صریحی برای بازیابی وجود ندارد، مدل همچنان به الگوهای مفهومی تثبیت‌شده در لایه‌های درونی خود تکیه می‌کند. برای مقابلهٔ مؤثرتر با این سوگیری، دو رویکرد دیگر که در سطح آموزش و بازتنظیم خود مدل مداخله می‌کنند، احتمال موفقیت بیشتری دارند: «آموزش نظارت‌شده» (Supervised Fine-Tuning, SFT) و «آموزش دامنه‌ای» (Domain-Adaptive Training, DAT). در روش نخست، مدل با مجموعه‌ای از پرسش‌وپاسخ‌های معتبر و سامان‌یافته آموزش می‌بیند تا الگوی پاسخ‌گویی‌اش با چارچوب معرفتی مشخصی هماهنگ شود (Ouyang et al., 2022, pp. 27731–27732)؛ در روش دوم

نیز مدل در معرض حجم وسیعی از متون یک حوزهٔ معرفتی خاص قرار می‌گیرد تا با زبان، مفاهیم و شیوه استدلال آن حوزه آشنایی عمیق‌تری پیدا کند (Gururangan et al., 2020, pp. 834–836). ازآنجاکه این دو روش در سطح پارامترها و الگوهای درونی مدل مداخله می‌کنند، ظرفیت بیشتری برای غلبه بر سوگیری سوم دارند. با این همه، این رویکردها نیز راه‌حل کاملی به شمار نمی‌آیند، زیرا بسیاری از چارچوب‌های هنجاری مدل‌ها در مراحل آغازین طراحی، پیش‌آموزش و سیاست‌های هم‌راستاسازی آن‌ها تثبیت می‌شود و تنظیم‌های بعدی معمولاً در محدودهٔ همان ساختار کلی عمل می‌کند. ازاین‌رو، هرچند ترکیب این روش‌ها می‌تواند شدت این نوع سوگیری را کاهش دهد، راهکار ریشه‌ای در بلندمدت توسعهٔ مدل‌های بومی است که از مرحلهٔ طراحی و پیش‌آموزش بر اساس مبانی معرفتی و ارزشی مدنظر ساخته شوند و نظام هم‌راستاسازی و ارزیابی آن‌ها نیز از آغاز بر همین اساس شکل گیرد. البته در کنار این تمهیدات فنی، آموزش کاربران نیز اهمیت دارد. آنان باید بدانند که مدل‌های زبانی به‌شدت از بافتار پرسش اثر می‌پذیرند؛ لذا باید از کلی‌گویی پرهیز کرده و قیود لازم را به‌روشنی ذکر کنند. برای مثال بنویسند: «پاسخ را صرفاً بر اساس فقه امامیه و منابع معتبر شیعی ارائه کن». این قیدگذاری، هرچند سوگیری‌های عمیق مدل را از میان نمی‌برد، می‌تواند احتمال فعال‌شدن چارچوب‌های عمومی یا ناسازگار را کاهش دهد.

پیچیدگی متون دینی

سومین عامل مهمی که کاربرد هوش مصنوعی را در آموزش دینی محدود ساخته و موجب بروز خطا در پاسخ‌های آن می‌شود، نه در داده و نه در الگوریتم، بل در پیچیدگی ذاتی متون دینی ریشه دارد. داستان این متون با بسیاری از متون دیگر، مانند متون علمی یا گزارش‌های خبری، کاملاً متفاوت است، زیرا ویژگی‌های متمایزی دارند که فهم آن‌ها را برای سیستم‌های ماشینی، که برای درک زبان صریح و تحت‌اللفظی طراحی شده‌اند، دشوار می‌سازد. این پیچیدگی، که ناشی از لایه‌های عمیق معنایی، فرهنگی و زبانی تنیده در این متون است، عمدتاً خود را در دو سطح اصلی نشان می‌دهد: نخست، در سطح معنا و دلالت‌های زبانی، و دوم، در سطح ساختار و منطق تفسیری حاکم بر مجموعه منابع. در ادامه این دو جنبهٔ اصلی را شرح می‌دهیم.

نخستین و ملموس‌ترین جنبه از این پیچیدگی، به ویژگی‌های خاص زبان دین مربوط می‌شود. قرآن و روایات، برخلاف متون خبری یا علمی، سرشار از ساختارهای زبانی‌ای هستند که به‌کمک مفاهیم انتزاعی (مانند «فطرت»، «تقوا» و «رضا» که صرفاً تعریف واژگانی ندارند)، استعاره‌های معرفتی (مانند «اللَّهُ نُورُ السَّمَاوَاتِ وَالْأَرْضِ» یا «ید الله فوق ایدیهم»)، تمثیل‌های داستانی (چون داستان موسی و خضر که ظاهر آن حامل لایه‌های باطنی است) و کنایه‌های ظریف، افقی فراتر از درک سطحی می‌گشایند. هدف این زبان انتقال صرف اطلاعات نیست، بلکه برانگیختن تأمل، ایجاد تحول درونی و زمینه‌سازی برای تجربهٔ معنوی در مخاطب است. تفسیر چنین متنی مستلزم توانایی عبور از ظاهر الفاظ، تشخیص سطوح معنایی و جای‌دهی هر واژه در منظومهٔ مفهومی دین است. در مقابل، مدل‌های هوش مصنوعی که بر الگوهای آماریِ واژگان تکیه دارند و فاقد تجربهٔ زیسته، شهود معنایی و درک بافت فرهنگی‌اند، ممکن است از درک مفهوم واقعی آن عاجز مانده و تفسیری تک‌بعدی و تحت‌اللفظی از آن ارائه می‌دهند. برای مثال، ممکن است برخلاف باورهای شیعی تفسیر

جسم‌انگارانه از مفهوم «وجه الله» ارائه دهند یا «قلب سلیم» را به یک اصطلاح روان‌شناختی صرف فرو بکاهند. این رویکرد، نه‌تنها مخاطب را از جوهر معرفتی و معنوی متن محروم می‌سازد، بل می‌تواند به شکل‌گیری برداشت‌های سطحی و انحرافی از مفاهیم عمیق دینی منجر شود.

دومین جنبه از پیچیدگی متون دینی، به ساختار کلی بدنهٔ معرفت دینی بازمی‌گردد که از منابع متعدد و متنوع تشکیل شده است. برخلاف برخی از حوزه‌های دانشی که ساختاری منسجم و خطی دارند، معرفت دینی برآمده از مجموعه‌ای وسیع از داده‌هاست: آیات قرآن، احادیث نبوی و روایات ائمه با درجات اعتبار متفاوت و تفاسیر گوناگون. ازآنجاکه این داده‌ها در ظاهر یا حتی در واقع بعضاً با هم تعارض دارند، فهم روشمند آن‌ها لاجرم باید در پرتو اصول اجتهادی، وزن‌دهی به منابع، ترجیح سند قوی بر ضعیف، تفکیک دلالت قطعی از ظنی، و تشخیص نص از ظاهر صورت بگیرد. این ویژگی ذاتی متن دینی، آن را از همان پایه به موضوعی نامتجانس با منطق مدل‌های زبانی بدل می‌سازد و آنچه را پیش‌تر در بخش سوگیری الگوریتمی پذیرش غیرانتقادی گفتیم، تشدید می‌کند: اگر الگوریتم حتی در مواجهه با متون هم‌سطح نیز توان غربال ندارد، در برخورد با این شبکهٔ سلسله‌مراتبی و متکثر، شکاف میان منطق ماشین و مقتضای فهم دینی عمیق‌تر می‌شود. به سخن دیگر، چندمنبعی بودن، هرچند به خود متن دینی مربوط است و از این‌رو عاملی متمایز از سوگیری مذکور به شمار می‌آید، دامنه و شدت پیامدهای آن سوگیری را افزایش می‌دهد؛ به‌گونه‌ای که یک خطای نسبتاً قابل‌مدیریت را به مسئله‌ای با ابعاد ساختاری بدل کرده و در نتیجه، کارآمدی هوش مصنوعی را به‌عنوان ابزاری قابل اعتماد برای انتقال مفاهیم دینی، به‌شدت زیر سؤال ببرد.

البته دشواری‌های برخاسته از پیچیدگی زبانی و چندمنبعی بودن متون دینی، حکم به طرد مطلق هوش مصنوعی در این حوزه نمی‌دهد، بل با نشان دادن موانع موجود، مسیر صحیح کاربست هوش مصنوعی را مشخص می‌سازد: حرکت از مدل‌های پاسخ‌گو که به دنبال ارائه جواب قطعی هستند، به‌سمت ابزارهای روشنگر که وظیفه‌شان، پرده‌برداری از همین پیچیدگی‌ها برای کاربر است. بنابراین، می‌توان امیدوار بود که با همکاری نزدیک میان متخصصان علوم کامپیوتر و محققان علوم اسلامی، ابزارهایی طراحی شوند که با بهره‌گیری از ساختارهای معنایی نظیر گراف‌های دانش، به‌جای ساده‌سازی تقلیل‌گرایانه، نقش نقشه‌بردار این چشم‌انداز پیچیده معرفتی را ایفا کنند. ابزاری که با عرضهٔ تفاسیر مختلف یک استعاره، یادآوری درجهٔ اعتبار یک روایت و فهرست کردن نظرگاه‌های رقیب، چشم‌انداز پیچیدهٔ معرفتی را برای متخصصان روشن‌تر سازد و دسترسی آن‌ها را به این لایه‌ها عمق بخشد. در کنار این تحول فنی، انتظارات کاربران نیز باید اصلاح شود. کاربر باید آگاه باشد که متون دینی به‌دلیل لایه‌های معنایی و استعاری، با متون صریح علمی یا گزارش‌های خبری تفاوت فاحش دارند. بنابراین، باید سطح انتظارات خود از دقت پاسخ‌های هوش مصنوعی در این حوزه را تعدیل کند و همواره احتمال عدم درک ظرایف متن توسط ماشین را در نظر داشته باشد.

نتیجه‌گیری

همان‌طور که تبیین کردیم، استفاده از مدل‌های هوش مصنوعی در حوزهٔ مفاهیم دینی، به‌رغم مزایای بسیار، با چالش جدی تحریف مفاهیم مواجه است. نوشتار حاضر نشان داد که برای فهم ریشه‌های این چالش، می‌توان از یک چارچوب تحلیلی سه‌وجهی بهره برد. این چارچوب که بر پایه کالبدشکافی فرآیند تولید پاسخ استوار است، سه عامل اصلی را به‌عنوان منشأ این خطاها معرفی می‌کند: نخست، سوگیری در داده‌ها که در بازنمایی ضعیف منابع دینی، اسلامی و به‌ویژه شیعی در پیکرهٔ آموزشی مدل‌ها ریشه دارد؛ دوم، سوگیری در الگوریتم‌ها که از پذیرش غیرانتقادی اطلاعات، هذیان‌گویی در غیاب داده، و اعمال خط‌قرمزهای ارزشی سازندگان ناشی می‌شود؛ و در نهایت، پیچیدگی ذاتی متون دینی که از ساختار شبکه‌ای و کل‌گرای گزاره‌های دینی سرچشمه می‌گیرد. این سه عامل در هم‌افزایی با یکدیگر، صلاحیت هوش مصنوعی را در حوزهٔ انتقال مفاهیم دینی به‌شدت زیر سؤال می‌برند. با این حال، باید توجه داشت که تأثیر چالش‌های مذکور بر همهٔ کاربردهای هوش مصنوعی همیشه به یک اندازه و به یک منوال نیست. بر این اساس، می‌توان کاربردهای هوش مصنوعی در حوزهٔ انتقال مفاهیم دینی را از حیث مخاطرهٔ معرفتی به سه دستهٔ متمایز تقسیم کرد:

(۱) کاربردهای محتوایی (حوزهٔ پرخطر) که در آن، هوش مصنوعی مستقلاً به ارائهٔ محتوای دینی (مانند پاسخ به شبهات، ارائهٔ مشاوره اعتقادی، یا استنباط حکم شرعی) می‌پردازد. این حوزه به‌دلیل درگیری مستقیم با هر سه چالش معرفتی، در بالاترین سطح خطر قرار دارد و استفاده از هوش مصنوعی در آن، مشروط به اعمال اصلاحات فنی همه‌جانبه (در هر دو سطحِ ساختار درونی و خروجی مدل) و نظارت مستمر کارشناسان دینی بر فرآیند اعتبارسنجی است.

(۲) کاربردهای شکلی (حوزهٔ امن) که در آن، هوش مصنوعی صرفاً به مدیریت و بهینه‌سازی فرآیندهای ساختاری و اجرایی (مانند سازماندهی منابع، قالب‌بندی متون، یا تحلیل رفتارهای کاربری) می‌پردازد. ازآنجاکه این کاربردها با محتوای دینی تماس مستقیم ندارند، از چالش‌های معرفتی مصون بوده و امن‌ترین بستر برای بهره‌گیری از هوش مصنوعی به‌شمار می‌روند.

(۳) کاربردهای دوگانه (حوزهٔ مشروط) که شامل فعالیت‌هایی با کارکرد منعطف، فعالیت‌هایی نظیر «استخراج اطلاعات متنی» یا «مأخذیابی و ارجاع‌دهی به منابع» را شامل می‌شوند. مرز میان امنیت و خطر در این کارکردها، به میزان مداخله هوش مصنوعی در تحلیلِ محتوا بستگی دارد؛ تا زمانی که عملکرد ماشین صرفاً به شناسایی مکانیکی کلمات یا یافتن نشانی دقیق یک مطلب در یک پایگاه داده مشخص محدود باشد، آسیب‌پذیری معرفتی ناچیز است. اما به‌محض آنکه از ماشین خواسته شود تا به «ارزیابی اعتبار راویان» یا «ترجیح و داوری میان منابع متضاد» بپردازد، به فعالیتی پرخطر تبدیل می‌شود که نیازمند همان ضوابط سخت‌گیرانه دسته اول است.

در پرتو این تفکیک، راهبرد نهایی برای استفاده از هوش مصنوعی در حوزهٔ مفاهیم دینی مستلزم اجرای دو اقدام موازی است که یکی حوزهٔ طراحی را هدف می‌گیرد و دیگری حوزهٔ کاربری را. در حوزهٔ طراحی، متخصصان و برنامه‌ریزان باید از یک‌سو روی توسعهٔ مدل‌های زبانی بومی سرمایه‌گذاری کنند و از دیگرسو، در

اقدامی اقتصادی‌تر، روی مهار مدل‌های قدرتمند موجود تمرکز نمایند و با به‌کارگیری روش‌های RAG و RLHF جهت رفع خطاهای فنی و فرآیندی، و ابزارهای SFT و DAT برای اصلاح سوگیری‌های ارزشی و دامنه‌ای، خطر تولید محتوای نادرست مدل‌های موجود را کاهش دهند. همزمان، در حوزهٔ کاربری، باید بر توانمندسازی افراد از طریق آموزش سواد دیجیتال دینی تمرکز شود. این توانایی به کاربران کمک می‌کند تا ضمن بهره‌برداریِ مطمئن از ظرفیت‌های شکلی، در مواجهه با کاربردهای دوگانه و محتوایی رویکردی هوشمندانه و محتاطانه داشته باشند. تنها با چنین تلاش مشترکی است که می‌توان اطمینان یافت معارف اصیل شیعی فدای جذابیت‌های سطحی یک ابزار ناکامل نمی‌شود و این فناوری، از منبعی بالقوه برای تحریف، به وسیله‌ای کارآمد برای تسهیل دسترسی و تعمیق فهم معارف دینی بدل می‌گردد.

منابع

احمدوند، عباس. (۱۳۷۷). گذری بر مطالعات شیعی در غرب. مقالات و بررسی‌ها، ۶۳، ۱۵۳-۱۸۳.

اسکندری، سهراب. (۱۴۰۴، اردیبهشت). تحلیل جامع سوگیری در هوش مصنوعی: از داده تا الگوریتم. سیمرغ هوش مصنوعی. بازیابی‌شده از https://simorghai.ir/

بقایی، حسین، کارآمد ثانی، امین، و احمدی، ناصر. (۱۴۰۳). کاربرد هوش مصنوعی در آموزش. در مجموعه مقالات همایش پژوهش‌های مدیریت و علوم انسانی در ایران، ۱۶، ۲۴۱۰-۲۴۲۳.

تیلور، چالز. (۱۴۰۰). عصر سکولار (ترجمه علیرضا پاک‌نژاد). تهران: پگاه روزگار نو.

خدادادی، سینا، نصرتی‌آذر، ثنا، و کاظم‌زاده، سینا. (۱۴۰۲). هوش مصنوعی و آموزش. در همایش پژوهش‌های مدیریت و علوم انسانی در ایران، ۱۲، ۱۲۳۶-۱۲۴۲.

خیامی، مهسان، طلوعی، مهدیه، و حدادکاشانی، نرجس. (۱۴۰۲). ادغام هوش مصنوعی در آموزش و یادگیری. مطالعات روان‌شناسی و علوم تربیتی، ۹۷(۵)، ۳۷۱-۳۸۸.

ذوالفقاری، مهناز. (۱۴۰۱). شناسایی کاربردهای هوش مصنوعی در توسعهٔ کارآفرینی آموزشی [پایان‌نامه کارشناسی ارشد، دانشگاه تهران].

ربیعی‌زاده، احمد. (۱۴۰۰). کاربرد هوش مصنوعی در پژوهش‌های علوم اسلامی. ره‌آورد نور، ۲۰(۷۵)، ۲۸-۳۷.

سهرابی، نازیلا. (۱۴۰۳). شناسایی مؤلفه‌های هوش مصنوعی و کاربرد آن در بهبود کیفیت نظام آموزش و پرورش [پایان‌نامه کارشناسی ارشد، دانشگاه سمنان].

مصطفوی، سیدمحمدعلی. (۱۴۰۱). امکان‌سنجی کاربرد هوش مصنوعی در برنامه‌ریزی درسی (آموزش عالی) [پایان‌نامه کارشناسی ارشد، دانشگاه شهید بهشتی].

نبوی، سیدمجید، آقابراری، زهرا، و نریمان، کیانوش. (۱۴۰۳). هوش مصنوعی و آموزش‌های دینی. اراک: سازمان انتشارات جهاد دانشگاهی واحد استان مرکزی.

Abubakar, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298–306). Association for Computing Machinery.

Andriani, A. D., & Sudirman, S. (2023). Cyberreligion: The role of artificial intelligence as a communication medium for religious education learning in the digital era. *TARBAWY: Indonesian Journal of Islamic Education*, *10*(2), 171–180.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). New York, NY: Association for Computing Machinery.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264–75278.

Digital Islamic Studies Curriculum. (n.d.). *Digital Islamic studies curriculum*. University of Michigan. Retrieved August 21, 2025, from https://sites.lsa.umich.edu/digitalislam/

Hutchinson, B. (2024). Modeling the sacred: Considerations when using religious texts in natural language processing. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 1029–1043). Association for Computational Linguistics.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (pp. 9459–9474). Neural Information Processing Systems Foundation.

Limna, P., Siripipatthanakul, B., Phayaphrom, P., & Siripipattanakul, S. (2022). A review of artificial intelligence (AI) in education during the digital era. *Advance Knowledge for Executives*, *1*(1), 1–9.

Luccioni, A., & Viviano, J. (2021, August). What's in the box? An analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 182–189). Association for Computational Linguistics.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 27730–27744). Neural Information Processing Systems Foundation.

Pew Research Center. (2009). *Mapping the global Muslim population: A report on the size and distribution of the world's Muslim population*. Washington, D.C.: Pew Research Center.

Pham, S. T. H., & Sampson, P. M. (2022). The development of artificial intelligence in education: A review in context. *Journal of Computer Assisted Learning*, *38*(5), 1408–1421.

Salsabila, A.-Z. A., & Rohiem, A. F. (2023). The ethical influence of artificial intelligence (AI) in religious education. In *Proceeding International Conference on Religion, Science and Education* (Vol. 3, pp. 81–88).

Saputra, I., Sriadhi, S., Mursid, R., & Saputra, D. A. (2022). Integration of artificial intelligence in education: Opportunities, challenges, threats, and obstacles. *The Indonesian Journal of Computer Science*, *12*(4), 1590–1600.

Wang, S., Liu, W., Atif, Y., & Fourtassi, A. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*, 124167.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., S., S., Brown, T., Hawkins, W., & Stepleton, T. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088