AI and Ethics (2026) 6:463
https://doi.org/10.1007/s43681-026-01254-5

RESEARCH

Who speaks when machines read scripture: boundary-making and interpretive authority in AI-mediated religious knowledge

Hishomudin Ahmad[1] · **Norfarhana Ahmad Ghafar**[1]

Received: 27 February 2026 / Accepted: 27 June 2026 / Published online: 13 August 2026
© The Author(s) 2026

Abstract
This article examines how large language models increasingly mediate access to religious and culturally normative knowledge, where interpretive claims carry institutional and moral authority rather than merely informational value. It develops a boundary-making account of AI governance that treats interpretive authority not as a function of representational accuracy but as an outcome of sociotechnical practices across the AI pipeline, from the formalization of sacred corpora into datasets to model alignment and platform circulation. Drawing on science and technology studies, the analysis shows how these transformations shape both what meanings become legible and who is authorized to speak in the name of the text. The article introduces the Minimum Reporting Standards for Digital Analysis of Sacred Texts (MRS-DST) as a governance framework centered on disclosure across textual, algorithmic, and institutional domains, repositioning AI as an assistive analytic instrument rather than an interpretive arbiter in religious knowledge systems.

Keywords Interpretive authority · AI governance · Digital hermeneutics · Science and technology studies · Sacred text analysis

1 Introduction

Large language models (LLMs) are increasingly consulted in domains where knowledge carries normative force alongside informational value, including religious interpretation, pedagogy, and public moral discourse. In such settings, the central risk is not limited to technical unreliability or model bias. It is a problem of governance and authority: when systems are asked to explain scripture, justify ethical claims, or translate doctrinal concepts into actionable guidance, they participate in shaping what counts as legitimate knowledge, who is authorized to speak, and under what institutional conditions interpretive claims acquire normative force [10, 13, 18]. Prevailing AI ethics frameworks typically respond through principles of transparency, fairness, and accountability, treating authority as a derivative of representational accuracy or procedural openness. While necessary, such approaches remain conceptually insufficient for religious and culturally grounded knowledge traditions in which legitimacy is historically institutionalized, ethically disciplined, and socially authorized rather than inferred from output correctness alone [4, 23, 24].

This study reframes AI-mediated interpretation as an institutional design problem rather than a technical performance problem. It asks not whether systems can be made objective or unbiased, but how interpretive authority is produced, transferred, and stabilized through the governance arrangements that surround AI deployment. Digital Religion studies has demonstrated how digital technologies reshape religious authority through changing communicative practices and technological mediation [8, 9]. However, it has paid less attention to how interpretive authority is constituted when AI participates in producing religious knowledge. This article addresses that gap by examining how AI-mediated interpretive authority is produced, validated, and governed. These arrangements include corpus selection and canon policy, annotation and prompting protocols, model constraints and disclosure regimes, and organizational settings, such as journals, educational institutions, platforms, and regulatory bodies, within which AI

✉ Hishomudin Ahmad
hishomudin@usim.edu.my

✉ Norfarhana Ahmad Ghafar
norfarhana@usim.edu.my

[1] Universiti Sains Islam Malaysia, Bandar Baru Nilai, Malaysia

outputs are reviewed, circulated, and sanctioned. Across the generative pipeline, successive transformations from cultural corpora to datasets, from datasets to base models, and from base models to instruction-tuned systems enact epistemic and normative boundaries that delimit what forms of meaning become legible, what interpretive options are foregrounded, and which traditions of validation remain visible or are displaced [11, 17, 22].

On this basis, the paper introduces the Minimum Reporting Standards for Digital Analysis of Sacred Texts (MRS-DST) as a governance mechanism oriented toward disclosure rather than output policing. Rather than positioning AI as an interpretive arbiter, MRS-DST operationalizes institutional accountability by requiring structured reporting of textual provenance, methodological traceability across human–machine divisions of labor, algorithmic mediation constraints, validation pathways anchored in recognized interpretive communities, and context-specific ethical oversight. The contribution is twofold: a conceptual account of interpretive authority as an infrastructural outcome of governance and boundary-making in AI pipelines, and a practical reporting standard that journals, educational institutions, and platforms can use to discipline the circulation of AI-generated religious knowledge without collapsing normative traditions into model-centric notions of accuracy [13, 18]. In doing so, the article advances the study of Digital Religion by explaining how interpretive authority is institutionally constituted when AI participates in the production, validation, and circulation of religious knowledge.

2 Authority, representation, and interpretive mediation

Understanding AI-mediated religious interpretation requires more than evaluating algorithmic performance or ensuring ethical compliance. It requires a theory of interpretive authority. Existing theories of authority, legitimacy, representation, and sociotechnical mediation have generated important insights across political theory, sociology, Science and Technology Studies (STS), and Digital Religion. However, these traditions have largely developed as parallel bodies of knowledge with distinct analytical priorities. Consequently, they explain why authority is recognized as legitimate, how representative claims are constituted, and how institutional practices stabilize knowledge, yet they do not adequately explain how these dimensions converge when AI participates in the interpretation of sacred texts. This theoretical fragmentation leaves an important conceptual gap. To address this gap, the following framework integrates these complementary perspectives into an integrated theory of interpretive authority, conceptualized as a relational, institutional, and sociotechnical phenomenon underpinning AI-mediated religious knowledge.

2.1 What is interpretive authority?

Authority is a foundational concept of social life. It explains why some individuals, institutions, and traditions are recognized as legitimate sources of judgment while others are not. The central issue is therefore not who possesses the greatest capacity to influence others, but whose judgments are accepted as rightful and why [37]. Addresses this distinction through the concept of *Herrschaft* (legitimate authority), arguing that authority exists only when those expected to comply recognize the legitimacy of the order itself. Obedience is therefore grounded less in coercion than in a shared belief that particular actors possess the legitimate right to exercise judgment. This distinction is especially significant in religious interpretation, where the acceptance of an interpretation depends not only on its content but also on the recognized legitimacy of the interpreter.

Authority should not be conflated with power [37]. Defines *Macht* (power) as the probability of imposing one's will despite resistance, whereas authority presupposes voluntary recognition and normative acceptance. The difference is fundamental. Governments may exercise regulatory power over religious institutions, digital platforms may shape public attention through algorithmic recommendation, and large language models may process extensive religious corpora, identify linguistic patterns, and generate sophisticated theological explanations within seconds. These capabilities demonstrate computational power rather than interpretive authority. Persuasive outputs may simulate legitimate interpretation, yet computational capability explains only what AI is able to do, not what it is institutionally entitled to do. This distinction provides the conceptual foundation for the argument developed throughout the present study.

Legitimacy provides the normative foundation upon which authority rests. Without legitimacy, authority is reduced to influence or compliance [37]. Argues that authority becomes stable only when a social order is recognized as valid by those who participate in it. Building on this insight, Jasanoff [18] demonstrates that societies establish culturally specific procedures through which knowledge claims acquire public credibility and institutional recognition. Accordingly, the authority of knowledge rests not on its substantive content alone but on the institutional arrangements governing its validation, including professional norms, peer review, and shared standards of evaluation. Religious interpretation follows the same logic. Interpretive claims become authoritative because they emerge from recognized scholarly traditions, pedagogical lineages, communal recognition, and

ethical accountability. Authority is therefore institutional before it is individual.

A further distinction concerns expertise. Although authority and expertise frequently overlap, they perform different functions. Expertise refers to specialised knowledge, technical competence, and disciplinary mastery acquired through education and experience. Authority concerns the socially recognized right to exercise that expertise on behalf of a community. Longino [22] argues that knowledge becomes authoritative not because experts possess superior information, but because their claims are subjected to organized criticism, communal evaluation, and institutional scrutiny. Expertise is therefore a necessary but insufficient condition for authority. This distinction becomes particularly important in the context of generative AI. Contemporary language models demonstrate remarkable competence in semantic retrieval, linguistic analysis, and large-scale knowledge synthesis. They can summarize classical commentaries, compare legal opinions, and identify interpretive patterns with unprecedented speed and consistency. These capabilities reflect computational expertise rather than interpretive legitimacy. AI systems neither inherit institutional mandates nor assume moral responsibility for the consequences of interpretation, nor do they participate in the communal processes through which religious authority is recognized, contested, and sustained.

The distinction between expertise and authority becomes even more consequential as religious interpretation is increasingly mediated through intelligent technologies. Contemporary AI ethics frameworks generally assume that improvements in accuracy, explainability, transparency, and fairness enhance the reliability of AI-generated interpretations [13]. More recent governance frameworks extend this emphasis by prioritizing accountability, contestability, and meaningful human oversight [26]. These developments undoubtedly strengthen computational reliability, but they leave unresolved the question of interpretive legitimacy. Legitimate authority depends upon institutional recognition rather than technical capability alone [37]. Likewise, authoritative knowledge emerges through organized critical scrutiny and communal validation rather than expertise alone [22]. An explanation may therefore be accurate without being authoritative, and a response may be transparent without possessing the recognized right to speak on behalf of a religious tradition. The legitimacy of interpretive claims is established through culturally embedded practices of public validation rather than technical correctness alone [18]. AI systems may consequently exhibit extraordinary analytical competence while lacking the institutional recognition required to guide doctrinal understanding. The challenge is therefore fundamentally normative rather than epistemic because interpretive legitimacy in religious traditions remains grounded in historically constituted institutions, ethical discipline, and communal authorization [4].

Taken together, these theoretical perspectives clarify why authority depends upon legitimacy, institutional recognition, and socially organized processes of validation. What they do not explain, however, is how these dimensions converge when AI actively participates in the production, validation, and circulation of religious knowledge. Building on Weber's theory of legitimate authority [37], Pitkin's [30] account of authorized representation, Saward's [32] theory of representative claims, and Science and Technology Studies [18, 21], this article conceptualizes interpretive authority as the institutionally authorized, socially recognized, and sociotechnically enacted capacity to produce, validate, and circulate interpretations on behalf of a community of meaning. This definition distinguishes interpretive authority from technical expertise, computational capability, and algorithmic performance. Rather than residing in individual actors, interpretive authority emerges through the convergence of institutional authorization, representative mediation, communal recognition, and accountable governance. This conceptualization provides the theoretical foundation for the next section, which examines how such authority is exercised through processes of representation and the right to speak on behalf of religious traditions.

2.2 Authority and representative mediation

If authority explains why particular interpretations are regarded as legitimate, representation explains who is entitled to exercise that authority. The two concepts are inseparable but analytically distinct. An individual may possess extensive knowledge without being recognized as a legitimate representative of a religious tradition, just as institutions may exercise interpretive authority despite differences in the expertise of their individual members. The central question is therefore not simply who can interpret scripture, but who may legitimately interpret on behalf of a community of belief. This distinction becomes increasingly significant as AI systems generate interpretations that resemble authoritative religious guidance.

Representation is fundamentally a relationship of authorization. Pitkin [30] argues that representatives derive legitimacy not because they possess superior competence but because they are recognized as having the legitimate right to act or speak on behalf of others. Authorization therefore precedes authority. Religious interpretation illustrates this principle clearly. Classical exegetes, jurists, and religious scholars do not derive interpretive authority solely from intellectual achievement. Their legitimacy is grounded in scholarly lineages, institutional appointments, pedagogical traditions, and communal recognition that authorize them to

interpret scripture on behalf of particular interpretive communities. Representation is therefore institutional before it is individual.

Authorization alone, however, does not permanently secure representative legitimacy. Saward [32] reconceptualizes representation as an ongoing process of representative claim-making rather than a fixed institutional status. Individuals and institutions continually advance claims to speak for others, yet these claims become authoritative only when they are recognized by relevant audiences. Representation is therefore dynamic rather than static. It is negotiated rather than assumed. Religious interpretation follows the same logic. Interpretive authority depends not only on institutional authorization but also on the continuing recognition of interpretive communities that accept, contest, and reaffirm representative claims over time.

This perspective provides an important lens for understanding AI-mediated religious interpretation. Contemporary large language models increasingly generate responses that resemble the work of religious scholars by synthesizing classical commentaries, comparing legal opinions, and presenting theological conclusions in a coherent and persuasive form. Although AI systems do not explicitly claim religious authority, their communicative style implicitly performs representative functions. Users frequently engage with these systems as though they were speaking on behalf of a religious tradition rather than processing textual information. From Saward's perspective, AI-generated interpretations may therefore be understood as representative claims whose legitimacy depends upon audience recognition and institutional validation rather than computational sophistication alone.

The emergence of AI-mediated interpretation consequently shifts attention from technological capability to representative legitimacy. AI systems cannot authorize themselves, nor can algorithmic performance establish the right to speak for scripture. Representative authority arises only when interpretive claims are embedded within institutional procedures that authorize, evaluate, and hold those claims accountable. Cheong [9] argues that religious authority in digitally mediated environments is communicatively constituted through the interactions of religious leaders, media technologies, institutions, and audiences, rather than residing in technological media themselves. Digital mediation therefore reshapes the communicative ecology of religious authority while preserving the institutional foundations upon which representative legitimacy depends.

The Digital Religion literature extends this institutional perspective by showing that technological mediation redistributes, rather than replaces, religious authority.Research in Digital Religion reinforces this conclusion. Campbell [8] demonstrates that digital media redistribute visibility, participation, and access to religious knowledge without displacing the institutional foundations of religious authority. Building on this insight, Atallah [5] argues that AI-assisted religious guidance is best understood as a problem of institutional mediation rather than technological substitution. Although algorithmic systems increasingly participate in the production and circulation of religious knowledge, their legitimacy continues to depend on established traditions of legal reasoning, scholarly validation, and ethical oversight. The central question, therefore, is not whether AI can generate interpretations, but whether those interpretations can legitimately represent the traditions and communities from which religious authority is derived.

AI systems therefore function as mediators of interpretation rather than holders of interpretive authority. They may organize, synthesize, and communicate religious knowledge, but they cannot authorize themselves to speak on behalf of a religious tradition. The right to interpret derives from institutional authorization, representative legitimacy, and communal recognition rather than computational capability. The central question is therefore not whether AI can generate religious interpretations, but who is legitimately authorized to speak for a community of belief. This question of representative authorization provides the point of departure for the next section.

2.3 Interpretive authority as sociotechnical practice

The preceding sections established that interpretive authority depends upon legitimacy and representative mediation. Yet these concepts alone do not explain how authority becomes institutionally established and sustained over time. Legitimate representatives do not exercise authority in isolation, nor does legitimacy emerge independently of the social arrangements within which interpretation takes place. The more fundamental question therefore concerns the processes through which interpretive authority is produced, stabilized, and maintained. Science and Technology Studies (STS) provides a useful response by shifting attention from authority as an individual attribute to authority as a sociotechnical accomplishment [18, 21]. From this perspective, authority is not simply possessed. It is continuously enacted through the interaction of people, institutions, technologies, and governance practices.

Understanding authority as a sociotechnical accomplishment first requires abandoning the assumption that knowledge originates from autonomous individuals. Latour's [21] actor-network theory argues that knowledge is produced through heterogeneous networks comprising human actors, institutions, technologies, texts, and material infrastructures. Scientific facts become authoritative because these networks stabilize particular claims while rendering

competing alternatives less visible. Religious interpretation is no exception. Sacred texts, exegetical traditions, educational institutions, digital databases, language models, editorial practices, and scholarly communities collectively shape which interpretations become credible, influential, and publicly recognized. Interpretive authority therefore emerges through networks of relations before it becomes associated with particular interpreters.

Networked production alone, however, cannot explain why some interpretations appear objective while others remain contested. Haraway's [16] concept of situated knowledges addresses this limitation by rejecting the ideal of detached neutrality. Every interpretation is embedded within particular historical, cultural, and institutional conditions that shape both its possibilities and its limitations. The same principle applies to AI-mediated interpretation. Outputs generated by large language models are conditioned by training data, model architectures, prompting strategies, and design decisions that influence what is emphasized, omitted, or rendered plausible. Transparency therefore functions as more than a technical requirement. It becomes an essential condition for interpretive accountability because it makes visible the situated assumptions underlying computational interpretation.

Situated knowledge nevertheless raises a further question: how do particular interpretations acquire public legitimacy? Longino's [22] social epistemology argues that epistemic authority emerges through organized criticism, peer evaluation, and communal scrutiny rather than through expertise alone. Knowledge acquires legitimacy by passing through institutional processes of review, debate, and revision. Religious scholarship has long operated according to comparable principles. Scholarly lineages, pedagogical traditions, collective learning, and communal validation transform individual expertise into publicly recognized interpretive authority. Expertise is therefore a necessary condition for interpretation, but it becomes authoritative only when embedded within institutions capable of sustaining critical evaluation and accountability.

The institutional character of authority also varies across societies. Communities develop different procedures for determining whose knowledge deserves trust and how competing claims should be evaluated. Jasanoff's [18] concept of civic epistemologies explains these culturally specific practices by demonstrating that public credibility depends upon institutional procedures of authorization, review, and validation rather than upon knowledge claims alone. Daston and Galison [11] reinforce this insight by showing that even scientific objectivity is historically produced through evolving professional norms, technical infrastructures, and institutional practices. Objectivity is therefore not an inherent property of representations. It is an institutional accomplishment. Interpretive authority should likewise be understood as emerging through stable procedures of validation rather than through the intrinsic qualities of interpreters or interpretations themselves.

Once authority is understood as a sociotechnical accomplishment, the implications for AI governance become considerably clearer. Floridi et al. [13] argue that AI should be treated as a sociotechnical system whose ethical consequences arise from the interaction of technologies, institutions, and human actors rather than from algorithms in isolation. Extending this perspective, Novelli et al. [26] contend that legitimate AI governance depends upon institutional mechanisms that ensure transparency, accountability, contestability, and meaningful human oversight. Governance is therefore not an external constraint imposed upon AI after deployment. It constitutes one of the institutional conditions through which AI-mediated interpretation can acquire public legitimacy. The central question consequently shifts from whether AI produces accurate outputs to whether the institutional arrangements surrounding those outputs sustain interpretive accountability and public trust.

Thus, interpretive authority is institutionally enacted rather than individually possessed. It emerges through sociotechnical practices of authorization, validation, critical scrutiny, and public accountability that connect interpreters, institutions, technologies, and interpretive communities. AI systems may participate in these processes by supporting the production and circulation of religious knowledge, but they do not independently generate the institutional conditions from which legitimate authority arises. Interpretive authority should therefore be understood as an institutional accomplishment enacted through accountable sociotechnical practices. This understanding provides the foundation for the integrated theory developed in the following section.

2.4 Religious authority in digital and AI-mediated contexts

The emergence of digital technologies has transformed not only how religious knowledge is communicated but also how religious authority is constituted, recognized, and contested. While traditional authority remains anchored in established institutions, scholarly lineages, and communal practices of validation, digital platforms increasingly shape the visibility, accessibility, and credibility of religious interpretations. Religious knowledge now circulates across complex digital ecosystems where platforms, algorithms, and networked publics participate in its production and dissemination. The central question is therefore not whether digital technologies replace religious authority, but how they reconfigure the institutional conditions through which interpretive legitimacy is produced.

Campbell [8] conceptualizes this transformation through the notion of *networked religion*, arguing that digital media actively reshape religious authority rather than merely transmitting it. Online environments redistribute participation, expand access to religious knowledge, and diversify interpretive voices while leaving institutional authority intact. Religious authority consequently becomes more interactive, publicly negotiated, and continuously contested. Digital mediation therefore broadens participation while simultaneously redistributing the visibility and credibility of religious interpretation without eliminating the institutional foundations upon which legitimacy ultimately depends.

Building on Campbell's account of networked religion, Cheong [9] shifts attention from digital participation to the communicative constitution of religious authority in digitally mediated environments. Rather than treating digital media as neutral channels of communication, she argues that religious authority is negotiated through interactions among religious leaders, media technologies, institutions, and audiences. Digital media therefore reshape the communicative ecology of religious authority without displacing the institutional and ethical responsibilities that remain fundamentally human. These insights provide an important foundation for understanding AI-mediated religious interpretation, where intelligent systems increasingly participate in the communication of religious knowledge but do not independently acquire interpretive authority.

Extending this discussion further, Atallah [5] argues that AI-assisted religious guidance should be understood primarily as a problem of institutional mediation rather than technological innovation. Drawing upon *uṣūl al-fiqh*, he demonstrates that the legitimacy of AI-generated religious guidance depends not only upon computational capability but also upon methodological rigor, scholarly accountability, and institutional oversight. The legitimacy of AI-assisted interpretation therefore depends less on technological sophistication than on the institutional arrangements that authorize, evaluate, and govern its use.

These developments have encouraged scholars to reconceptualize religious authority as a distributed sociotechnical practice. Studies of digital Islamic authority argue that authority increasingly emerges through the interaction of textual traditions, technological infrastructures, communicative performance, and audience participation rather than residing exclusively within recognized scholars or formal institutions. Wahid and Abdulloh [39], for example, demonstrate that digital religious authority is negotiated through the combined influence of textual evidence, affective engagement, algorithmic visibility, and participatory interaction. Institutional traditions therefore remain indispensable, but they now coexist with new mechanisms of credibility generated by digital platforms.

Within this evolving landscape, it becomes necessary to distinguish algorithmic authority from interpretive authority. Algorithmic authority refers to the credibility users attribute to computational systems because they appear objective, efficient, consistent, or technologically sophisticated. Recent empirical research indicates that users increasingly attribute epistemic authority to generative AI despite limited awareness of its methodological constraints, uncertainty, and embedded biases [25]. Algorithmic authority therefore emerges from technological performance and user perception. Interpretive authority, by contrast, derives from institutional authorization, representative legitimacy, and communal recognition. The two concepts should not be conflated. AI systems may acquire considerable algorithmic credibility while lacking the normative authority required to interpret scripture on behalf of a religious community.

The distinction between algorithmic authority and interpretive authority provides the conceptual basis for differentiating AI as *musāʿid* and AI as *muḥakkim*. As *musāʿid*, AI functions as an assistive interpretive instrument that strengthens scholarly inquiry through linguistic analysis, corpus exploration, information retrieval, and comparative synthesis. Human scholars, institutions, and interpretive communities retain responsibility for validation, contextualization, and ethical judgment. As *muḥakkim*, however, AI assumes the position of interpretive arbiter by implicitly determining which interpretations become authoritative, which sources receive prominence, and which theological positions gain visibility. Such a shift transfers representative legitimacy from accountable interpretive communities to opaque computational infrastructures that neither possess institutional mandates nor assume moral responsibility for religious interpretation. The central ethical question is therefore not whether AI can generate interpretations of sacred texts, but whether algorithmic systems should ever occupy positions of interpretive authority traditionally reserved for institutionally authorized representatives.

These developments demonstrate that AI transforms the ecology of religious authority rather than replacing it [28]. Digital platforms, algorithmic systems, and generative AI increasingly participate in the production and circulation of religious knowledge, yet interpretive legitimacy continues to depend upon institutional authorization, representative mediation, communal recognition, and accountable governance [13, 18, 32, 37]. Existing theories illuminate important dimensions of this transformation but stop short of integrating them into a coherent explanation of AI-mediated interpretive authority. This article addresses that theoretical gap by conceptualizing interpretive authority as the institutionally authorized, socially recognized, and sociotechnically enacted capacity to produce, validate, and circulate

interpretations on behalf of a community of meaning. Figure 1 synthesizes this integrated theoretical model.

Figure 2 operationalizes the theoretical model presented in Fig. 1 by translating interpretive authority into an institutional governance framework for AI-mediated religious interpretation.

Figure 2 translates the theoretical model of interpretive authority into an operational governance architecture comprising five interrelated stages. Normative principles establish the ethical and epistemic foundations for AI-mediated interpretation. Institutional governance specifies the processes through which interpretive activities are authorized, supervised, and held accountable. MRS-DST reporting standards define the minimum documentation required to ensure transparency, traceability, and methodological accountability. Responsible deployment governs the communication, implementation, and ongoing monitoring of AI-generated interpretations. Finally, adaptive governance supports continuous evaluation, institutional learning, and periodic framework revision. Across all stages, human oversight, institutional accountability, transparency, and auditability operate as cross-cutting governance principles that sustain legitimate, accountable, and contextually grounded AI-mediated interpretive authority.

Building on these perspectives, this article extends Digital Religion scholarship by explaining how interpretive authority is institutionally constituted when AI participates in the production, validation, and circulation of religious knowledge. Rather than examining technological mediation alone, it integrates Weber's theory of legitimate authority [37], Pitkin's theory of political representation [30], Saward's theory of representative claims [32], Science and Technology Studies [11, 18], and AI governance scholarship [13, 26] into a unified theory of AI-mediated interpretive authority. The MRS-DST framework operationalizes this

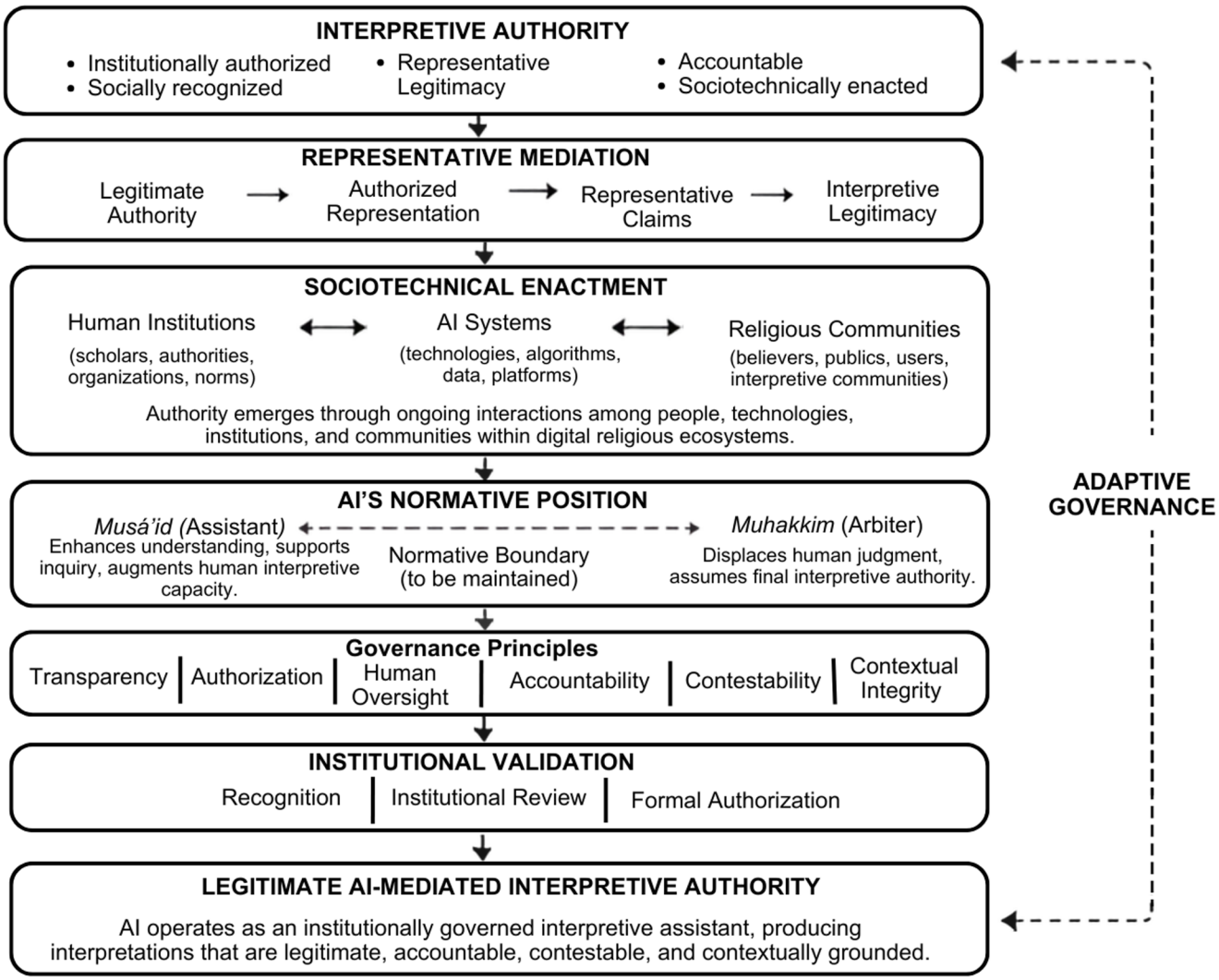


Fig. 1 Theoretical model of AI-mediated interpretive authority

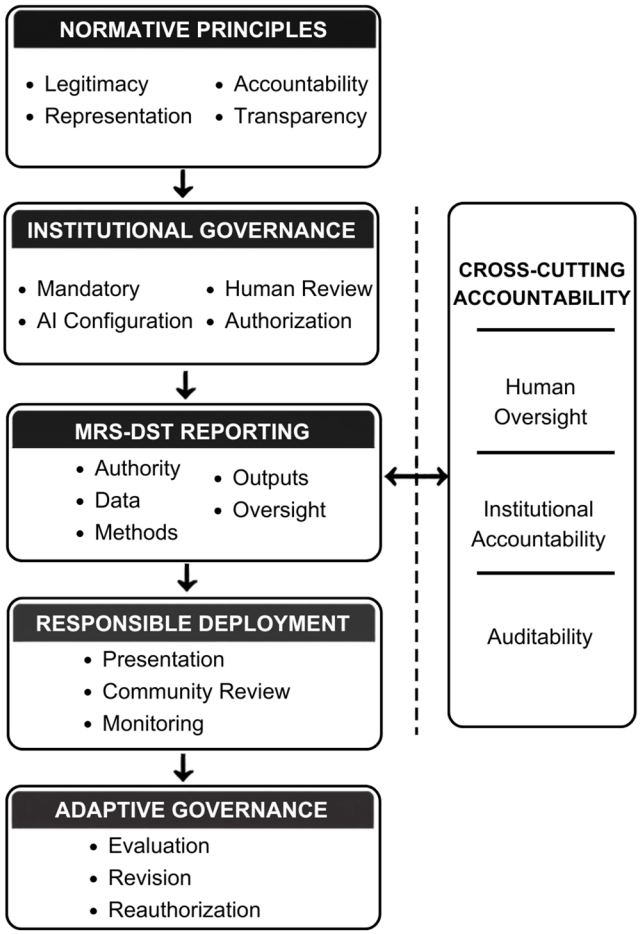


Fig. 2 MRS-DST governance framework

theoretical synthesis as an institutional governance architecture for accountable AI-assisted religious interpretation.

3 Authority boundaries in the AI pipeline

This study conceptualizes interpretive authority not as a property held by individual experts or technical systems, but as an institutional and procedural accomplishment that emerges through formal arrangements governing how knowledge is produced, validated, and circulated. Rather than treating authority as a derivative of representational accuracy or technical reliability, we approach it as a condition of legitimacy sustained through disclosure norms, review structures, and recognized communities of interpretation. Drawing on science and technology studies (STS), this study treats authority as a product of institutional design rather than representational quality, locating legitimacy in disclosure obligations, review procedures, and formally recognized validation communities that stabilize trust, accountability, and epistemic standing over time [11, 18, 22]. In AI-mediated religious knowledge systems, authority is enacted through policies, interfaces, datasets, and validation infrastructures that structure who may interpret, under what conditions, and with what normative force, rather than located in either human interpreters or machine outputs alone [10, 13].

3.1 Interpretive authority as institutional practice

Studies of knowledge governance has consistently shown that epistemic authority is sustained through institutional arrangements, accreditation structures, and procedural norms rather than concentrated in individual expertise alone. Latour's [21] actor-network theory and Pickering's [29] "mangle of practice" describe knowledge as the outcome of reciprocal adjustments between material systems and organizational constraints. In normatively grounded domains such as religion, however, these networks also perform regulatory functions. Authority is exercised through structured practices of accreditation, peer recognition, pedagogical transmission, and communal endorsement that determine which interpretations are recognized as legitimate and which are excluded.

Recent work on AI governance emphasizes that legitimacy in automated knowledge systems depends on institutional mechanisms of answerability, contestability, and limitation of power, particularly in settings where interpretive claims carry pedagogical, legal, or moral consequences [26, 36]. From this perspective, interpretive authority is not produced by systems in isolation, but by the institutional contexts that authorize their use, delimit their scope, and assign responsibility for their outputs. This view aligns with Jasanoff's [18] concept of "civic epistemologies," which frames knowledge legitimacy as a culturally embedded mode of governance. Applied to AI-mediated interpretation, authority becomes a property of configured relationships between platforms, scholarly communities, regulatory bodies, and users. What matters is not simply what a system can generate, but how its outputs are positioned within recognized structures of review, endorsement, and pedagogical or juridical consequence.

3.2 From objectivity to authority

This study enters into dialogue with STS analyses that foreground institutional practices as central to the production of objectivity. Daston and Galison's [11] historical account shows how epistemic virtues are stabilized through norms of documentation, visualization, and professional conduct rather than discovered as universal properties of representations. Recent analyses of AI governance demonstrate that

validation protocols, institutional review practices, and organizational policies function as formal sites where interpretive responsibility is assigned, constrained, and made publicly accountable [27].

We extend this line of argument by shifting the analytical focus from objectivity to interpretive authority. Rather than asking how systems produce true or unbiased representations, we ask how legitimacy, voice, and interpretive right are institutionally conferred and maintained. This shift reframes AI ethics as a problem of governance design: the central question becomes how platforms, journals, educational institutions, and policy frameworks structure the conditions under which AI-mediated interpretations are disclosed, attributed, reviewed, and sanctioned.

This pivot is particularly salient in religious knowledge systems, where authority is inherently normative and morally constituted [4]. Religious interpretation organizes ethical obligation, shapes communal identity, and guides social conduct rather than functioning solely as textual explanation. Accordingly, the authority to interpret is inseparable from processes of ethical formation and scholarly discipline [23], as well as institutional recognition and communal authorization [24]. When AI systems generate scriptural explanations, they extend beyond providing alternative readings by entering established moral and institutional economies of recognition, with the potential to redistribute epistemic standing among scholars, religious institutions, digital platforms, and users.

3.3 Cultural and religious epistemologies as governance frameworks (Asian perspective)

Asian religious traditions offer a distinctive model of authority as a form of governance rather than a purely methodological concern. In Islamic epistemology, interpretive practices such as *tafsīr, syarḥ*, and *sanad–matn* reasoning function as institutionalized systems of validation that embed interpretation within chains of transmission, ethical discipline (*adab*), and communal accountability [15, 31, 3]. Authority is enacted through scholarly lineages, pedagogical institutions, and normative conduct, not through interpretive autonomy or individual judgment alone.

This structure resonates with contemporary governance frameworks that emphasize human-centered oversight and contextual accountability. Contemporary governance frameworks emphasize that institutional design—through platform policy, journal standards, and educational regulations shape the moral and epistemic roles assigned to technical systems in public and scholarly life [13, 19]. Sun and Xu's [38] concept of "hybrid cognitive authority" similarly captures how knowledge emerges through structured collaboration between human and algorithmic agents within formal organizational settings.

Within this study, this alignment is formalized through the distinction between AI as *musāʿid* (assistive agent) and AI as *muḥakkim* (arbiter). This distinction reframes AI ethics as a problem of institutional architecture rather than technical optimization. It directs attention to how disclosure regimes, platform policies, journal standards, and educational governance structures determine whether AI-mediated interpretations are presented as provisional analytic signals or as de facto authoritative judgments.

These perspectives ground the theoretical claim that interpretive authority in AI-mediated religious knowledge is neither human-centered nor machine-centered but institutionally enacted through governance arrangements that organize validation, accountability, and circulation. The following sections develop reporting and disclosure mechanisms that treat authority as a designable and auditable outcome of institutional practice rather than an automatic by-product of computational capability.

4 MRS-DST: a governance framework

This section introduces MRS-DST as a governance framework for organising interpretive authority in AI-mediated sacred text analysis. It converts the article's central claim into a disclosure-based reporting structure that spans corpus formation, methodological design, model mediation, validation processes, and ethical oversight. The framework requires explicit documentation of how interpretive claims are produced, reviewed, and authorised across human and machine contributions. The following subsections set out the rationale for such standards, define the five governance domains, and present a reviewer-facing checklist for institutional use.

4.1 Rationale for standards: from optimization to disclosure

Current AI ethics and governance frameworks predominantly frame accountability as a problem of technical optimization, emphasizing bias reduction, performance metrics, and model-level transparency. While such interventions are necessary, they remain insufficient in domains where knowledge legitimacy is culturally, normatively, and institutionally grounded, such as religious interpretation. In these settings, the central governance question is not whether an AI system performs "well," but how interpretive authority is produced, transferred, and stabilized across institutional arrangements.

Research in science and technology studies (STS) has demonstrated that epistemic trust emerges from institutionalized practices of disclosure rather than from technical performance alone. Daston and Galison [11] characterize objectivity as a "moralized epistemology" sustained through shared norms of documentation, attribution, and procedural visibility, while Longino [22] locates reliability in structured critical interaction within recognized communities. Applied to AI-mediated religious knowledge systems, these insights imply that governance must prioritize institutional legibility of interpretive processes rather than model-level correctness.

This shift aligns with contemporary critiques of solutionist AI ethics, which caution against reducing sociotechnical problems to engineering challenges [33]. Governance therefore operates not only through technical design but through institutional sites such as journals, institutional review boards, universities, and platform policies where legitimacy is negotiated and formalized [13]. Within this institutional ecology, MRS-DST is proposed as a procedural infrastructure that formalizes disclosure obligations across the full AI pipeline and transforms reporting from an administrative requirement into a governance mechanism for interpretive authority.

4.2 The five governance domains

Figure 1 maps the five governance domains of MRS-DST (Table 1) onto successive stages of the AI interpretation pipeline, illustrating how interpretive authority is enacted through institutional and infrastructural arrangements rather than derived from model performance or output correctness alone.

4.2.1 Textual authority

This domain governs how sacred texts enter computational systems. Decisions concerning corpus provenance, canon

Table 1 Excerpt from a completed MRS-DST checklist (Qur'anic stylistics case)

Domain	Governance question	Required disclosure (completed)	Evidence/artifact	Reviewer action	Risk if omitted
1. Textual authority	Which Qur'anic corpus and exegetical tradition anchors interpretive claims?	Canonical corpus based on *Ḥafṣ ʿan ʿĀṣim* reading; verse segmentation aligned with Tanzil Uthmani standard; transliteration follows IJMES guidelines. Exegetical anchoring in al-Ṭabarī (*Jāmiʿ al-Bayān*) and Ibn ʿĀshūr (*al-Taḥrīr wa al-Tanwīr*). Variant readings documented but excluded from computational clustering	Corpus specification sheet; edition metadata; citation log of tafsīr sources; segmentation protocol (Appendix A)	Verify consistency between cited tafsīr sources and stylistic interpretations reported in Results. Confirm segmentation standard matches computational pipeline	Model outputs may be misconstrued as independent exegetical claims, displacing recognized scholarly authority
2. Methodological traceability	Where does human judgment end and algorithmic processing begin?	Human scholars define stylistic categories (parallelism, ellipsis, semantic compression). Model performs clustering and frequency mapping only. Theological interpretation remains exclusively human-authored	Annotation schema; prompt templates; workflow diagram (Fig. 2); authorship attribution statement	Confirm that interpretive claims are not generated by the model. Audit role separation in Methods section	Responsibility for interpretation becomes ambiguous, enabling attribution of theological claims to automated systems
3. Algorithmic mediation	What model constraints shape stylistic legibility?	Tokenization based on Arabic sub word segmentation (Sentence Piece). Embeddings: multilingual transformer (frozen weights). Prompt constraints prohibit semantic inference or normative evaluation. Known failure modes in handling elliptical constructions documented	Model card; prompt logs; limitation statement; error analysis appendix	Assess whether reported stylistic patterns fall within documented model capabilities	Computational artefacts may be elevated to theological or semantic judgments beyond model scope
4. Interpretive accountability	Who validates and authorizes interpretive conclusions?	Peer review by two Qur'anic studies scholars; internal review by institutional ethics committee for AI-assisted research; curricular use approved by department syllabus board	Reviewer reports; ethics approval letter; curriculum endorsement memo	Confirm presence of recognized validation bodies and review trail	Interpretive authority becomes platform- or author-centric rather than institutionally grounded
5. Ethical oversight	How are findings framed for non-specialist audiences?	Public-facing materials include disclosure banner: "AI-assisted stylistic analysis; theological interpretation by human scholars only." Educational deployment includes instructor guidance notes	Platform screenshots; disclosure template; teaching guide	Verify visibility of disclosure in published and pedagogical outputs	Users may treat AI-generated patterns as doctrinal or juridical authority

Illustrative audit extract demonstrating how interpretive authority is documented, reviewed, and institutionally validated under MRS-DST

selection, variant management, transliteration, and annotation constitute foundational institutional choices that delimit the interpretive space available to both models and users. Bowker and Star [6] demonstrate that systems of classification and standardization embed institutional values and power relations within technical infrastructures. Religious traditions have long established comparable governance mechanisms through *tafsīr* hierarchies, *sanad–matn* evaluation, and *adab al-ikhtilāf*, which regulate who is authorized to interpret and under what conditions [1, 7]. Accordingly, MRS-DST requires explicit disclosure of how digital corpora relate to these inherited epistemic authorities, ensuring that computational representations remain accountable to, rather than inadvertently displace, established traditions of validation (Fig. 3).

4.2.2 Methodological traceability

Interpretation in AI-mediated systems is distributed across human scholars, annotation protocols, prompting strategies, and automated processes. This domain therefore requires systematic documentation of annotation schemes, interpretive layers, and the division of labor between human and machine agents. Latour's [20] concept of *inscription devices* demonstrates that methods actively shape what becomes legible as knowledge rather than serving solely as mechanisms for recording phenomena. By rendering interpretive pathways traceable, MRS-DST positions methodological transparency as a prerequisite for institutional legitimacy rather than a secondary technical consideration.

4.2.3 Algorithmic mediation

This domain addresses how model architecture, tokenization schemes, embeddings, fine-tuning regimes, and prompting practices function as epistemic filters. These design choices shape what kinds of meanings can be expressed, recognized, or excluded before any interpretive engagement occurs. Critical algorithm studies emphasize that such systems operate as cultural and institutional mediators rather than neutral instruments [2, 14]. MRS-DST therefore requires authors to specify model limitations, training assumptions, and interaction constraints as part of their interpretive claims, situating algorithmic outputs within their institutional and material conditions of production.

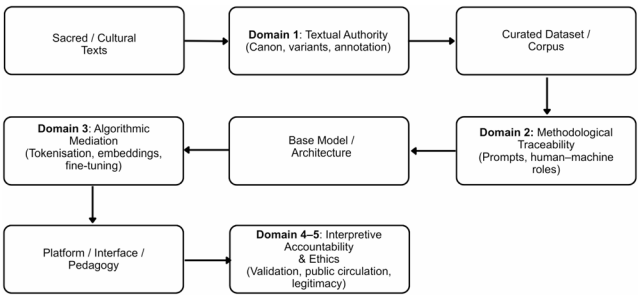


Fig. 3 Authority boundary-making across the AI interpretation pipeline

4.2.4 Interpretive accountability

In religious knowledge systems, authority is sustained through recognized communities of validation, such as scholarly peer review, juridical councils, or pedagogical institutions. This domain requires authors to identify who reviews, endorses, or contests AI-mediated interpretations and under what institutional frameworks such judgments are made. Longino's [22] model of epistemic communities and Serres's [35] notion of quasi-objects are relevant here: interpretive artefacts circulate within collectives, generating authority through structured participation rather than individual assertion. MRS-DST formalizes this circulation by making validation pathways visible and auditable.

Figure 1 maps the five governance domains of MRS-DST (Table 1) onto successive stages of the AI interpretation pipeline, illustrating how interpretive authority is enacted through infrastructural boundary-making rather than derived from model performance or output correctness alone.

4.2.5 Ethical oversight

The final domain addresses the broader social consequences of AI-mediated religious knowledge, including cultural sensitivity, pedagogical risk, and public dissemination impact. Floridi et al. [13] frames such concerns in terms of ethical governance ecosystems in which responsibility is distributed across developers, institutions, and users. In religious education contexts, related concerns include data privacy, normative influence, and the reshaping of pastoral authority. MRS-DST integrates these considerations into its reporting standards, ensuring that ethical evaluation extends beyond formal compliance into the domain of public trust and institutional legitimacy.

4.3 Reviewer-facing checklist: institutionalizing interpretive governance

To operationalize these domains, MRS-DST introduces a structured, reviewer-facing checklist (Table 1) designed for use by journals, ethics committees, and academic institutions. The checklist functions as a boundary object [37], enabling coordination between technical reviewers, domain scholars, and institutional gatekeepers without requiring epistemic consensus. It standardizes expectations for

disclosure across disciplines while preserving space for interpretive plurality.

Rather than evaluating the "correctness" of AI-generated interpretations, the checklist audits the institutional conditions under which interpretive claims are produced, validated, and circulated. In doing so, it repositions peer review and institutional oversight as active sites of governance, aligning digital sacred text research with established traditions of scholarly accountability. MRS-DST thus transforms reporting into a procedural architecture for sustaining interpretive authority in AI-mediated knowledge systems, ensuring that machines remain assistive agents rather than unexamined arbiters of religious meaning.

5 Operationalizing authority

This section presents two illustrative applications of the Minimum Reporting Standards for Digital Analysis of Sacred Texts (MRS-DST) to demonstrate how the framework operationalizes interpretive governance across the full AI pipeline. These applications are presented as conceptual illustrations of the proposed governance framework rather than as empirical validation, demonstrating how MRS-DST may be operationalized across different forms of AI-assisted religious research. The purpose of these applications is not to assess the empirical "accuracy" or theological correctness of AI outputs, but to show how institutional responsibility, disclosure obligations, and validation pathways can be made inspectable, auditable, and accountable at textual, methodological, algorithmic, and organizational levels. In this sense, the cases function as illustrative governance scenarios rather than empirical validation studies, demonstrating how interpretive authority can be documented, reviewed, and institutionally governed under MRS-DST.

5.1 Qur'anic stylistic analysis: delimiting exegetical authority and algorithmic pattern detection

The first application applies MRS-DST to a digital stylistic study of the Qur'an that integrates computational pattern detection with classical *tafsīr*. Under the Textual Authority domain, the reporting protocol requires explicit disclosure of corpus provenance, including canonical *qirāʾāt* selection, verse segmentation rules, and transliteration standards, alongside identification of the exegetical traditions used as interpretive anchors (e.g., *al-Ṭabarī, Ibn ʿĀshūr*). These disclosures function as institutional markers of legitimacy by clarifying how digital corpora are positioned in relation to recognized traditions of scholarly authority rather than presented as autonomous sources of interpretation [11].

Within the Methodological Traceability domain, the application documents the formal division of labor between human scholars and automated processes, specifying which stages of stylistic categorization, annotation, and interpretive framing remain under human judgment and which are delegated to algorithmic clustering or pattern extraction. Prompt templates, segmentation rules, and annotation schemas are reported as part of the study's procedural record, enabling reviewers and institutional gatekeepers to trace how analytic claims move from text to output and where interpretive discretion is exercised [20, 37].

The Algorithmic Mediation domain requires disclosure of tokenization schemes, embedding models, and interaction constraints used in identifying rhetorical features such as parallelism or ellipsis. These elements are framed as institutional design choices that shape the scope of computational legibility. By documenting these constraints, the standard prevents stylistic signals generated by the model from being implicitly elevated into exegetical or theological judgments [13].

Under Interpretive Accountability, the application specifies the institutional settings in which AI-assisted analyses are reviewed and endorsed, such as journal peer review, scholarly consultation, or curricular approval processes. This disclosure maintains a procedural separation between computational pattern detection and religious or semantic authority. The framework thus stabilizes AI's role as *musāʿid* (assistive analytic instrument) rather than *muḥakkim* (interpretive arbiter), anchoring legitimacy within recognized communities of validation rather than within the system itself [15, 24].

To demonstrate how MRS-DST operates as an institutional audit instrument rather than a conceptual framework alone, the following excerpts present completed reviewer checklists from two illustrative cases. The boxes are formatted as submission-ready artefacts designed for use by journals, ethics committees, and academic institutions. They show how disclosure, validation pathways, and role separation are documented in practice rather than inferred from narrative description.

5.2 *Ḥadīth* variant stability: tracing authority across transmission, corpora, and model outputs

The second application examines a study of ḥadīth variant stability that combines classical ʿilm al-ḥadīth with model-assisted textual comparison. Under the Textual Authority domain, MRS-DST requires authors to disclose the source collections used (e.g., *Ṣaḥīḥ al-Bukhārī*, *Ṣaḥīḥ Muslim*, and selected *Sunan* collections), the editorial editions consulted, and the criteria for variant inclusion. These disclosures render explicit the institutional and scholarly judgments that

define the legitimate object of analysis in both transmission-based *ḥadīth* scholarship and digital corpora. As Bowker and Star [6] demonstrate, systems of classification and standardization are never neutral,they embed institutional values and shape what becomes available for knowledge production.

Within the Methodological Traceability domain, the application reports how similar metrics, clustering thresholds, and human cross-checking procedures are combined. This documentation clarifies where probabilistic pattern recognition ends and interpretive judgment begins, ensuring that responsibility for evaluative claims remains anchored in recognized scholarly practice rather than attributed to automated processes. This separation reflects Longino's [22] model of epistemic communities, in which reliability emerges through structured critical interaction.

The Algorithmic Mediation domain requires disclosure of model limitations, training assumptions, and known failure modes in handling classical Arabic morphology and *isnād* structures. These disclosures position computational outputs as diagnostic indicators rather than as authenticity verdicts, preserving the normative priority of transmission-based evaluation and juridical reasoning.

Under Ethical Oversight, the application documents how findings are contextualized when circulated beyond specialist audiences, particularly in educational or public-facing environments. Disclosure statements, pedagogical guidance, and platform-level framing are used to prevent algorithmic outputs from being misconstrued as juridical or theological determinations. This emphasis aligns with governance approaches that locate accountability in institutional embedding, disclosure regimes, and public-facing design rather than in technical performance alone [13, 18] (Table 2).

Together, these audit excerpts illustrate how interpretive authority can be rendered institutionally visible and reviewable, providing a practical basis for the journal, platform, and policy implications discussed in Sect. 6.

Table 2 Excerpt from a completed MRS-DST checklist (Ḥadīth variant stability case)

Domain	Governance question	Required disclosure (completed)	Evidence/artifact	Reviewer action	Risk if omitted
1. Textual authority	Which ḥadīth corpora and editorial traditions anchor variant analysis?	Primary sources: *Ṣaḥīḥ al-Bukhārī* (Dār Ṭawq al-Najāh ed.), *Ṣaḥīḥ Muslim* (Dār Iḥyāʾ al-Turāth ed.), and selected *Sunan* collections. Variant inclusion criteria limited to isnād-linked textual divergences recorded in classical commentaries. Non-canonical digital variants exclude from clustering	Corpus provenance sheet; edition metadata; source justification memo (Appendix B)	Verify that all reported variants are traceable to recognized classical sources and editions	Computational corpora may supersede transmission-based scholarly authority
2. Methodological traceability	How are probabilistic similarity measures separated from scholarly judgment?	Model computes lexical and structural similarity scores only. Human scholars conduct *isnād* evaluation and *matn* assessment using classical criteria. Authenticity classifications are human-authored	Workflow chart; similarity threshold table; authorship attribution statement; review protocol	Confirm that authenticity determinations are not automated. Audit separation of analytic and evaluative stages	Model outputs may be misinterpreted as juridical verdicts on authenticity
3. Algorithmic mediation	What technical constraints shape detection of variant patterns?	Arabic morphological analyzer is integrated for token normalization. Known failure modes in handling weak narrators' names and *isnād* segmentation documented. Embeddings not trained on religious corpora	Model card; error log; morphology handling note; limitation appendix	Assess whether variant clusters fall within documented technical scope	Technical artefacts may be mistaken for authoritative *isnād* analysis
4. Interpretive accountability	Which institutions validate transmission-based conclusions?	External review by two specialists in *ʿilm al-ḥadīth*; internal faculty research ethics panel approval for AI-assisted textual analysis. Conference presentation vetted by disciplinary panel	Reviewer reports; ethics clearance; conference acceptance letter	Verify presence of recognized disciplinary validation trail	Authority shifts from scholarly community to platform or authors
5. Ethical oversight	How are results communicated to public and educational audiences?	Public summaries include disclaimer: "Computational analysis supports, but does not replace, classical *ḥadīth* authentication methods." Teaching materials include guidance on limits of model-assisted comparison	Disclosure banner: syllabus insert; outreach material samples	Check visibility of disclaimers and pedagogical framing	Readers may treat computational similarity as proof of authenticity

Illustrative audit extract demonstrating how interpretive authority is documented, validated, and institutionally governed in model-assisted transmission analysis

6 Institutional mediation

This study advances a shift from treating AI in religious knowledge systems as a problem of technical performance to understanding it as a site of institutional governance. MRS-DST is not proposed as a compliance checklist for ethical "best practice," but as a procedural architecture through which interpretive authority is rendered visible, contestable, and institutionally accountable. Rather than locating legitimacy in model behavior or output quality, the framework situates it in publicly inspectable practices of disclosure, review, and validation that organize how interpretive claims are produced, circulated, and authorized [11, 22]. By requiring structured reporting across textual, methodological, algorithmic, and organizational layers, MRS-DST reframes AI-mediated interpretation as a governed institutional process rather than a technical artefact awaiting ethical remediation.

6.1 Implications for journals: from method review to authority review

For academic publishing, MRS-DST extends peer review beyond methodological rigor toward authority auditing. Conventional review practices assess theoretical contribution, evidentiary sufficiency, and technical validity, yet often leave unexamined how datasets are canonized, how model design filters interpretive space, and how responsibility for claims is distributed between human scholars and automated systems. The proposed checklist introduces a formal mechanism for evaluating epistemic provenance, disclosure adequacy, and interpretive accountability, aligning with governance-oriented approaches to transparency in AI research [26]. Within this model, journals function as institutional gatekeepers of disclosure norms, shaping what counts as legitimate AI-mediated knowledge by demanding traceability, role separation, and validation pathways rather than performance optimization alone.

6.2 Implications for education: AI literacy as epistemic formation

Within religious studies and digital humanities curricula, MRS-DST reframes AI literacy as a matter of epistemic formation rather than technical competence. Students and educators are encouraged to examine how datasets, models, and platforms reorganize relations of authority between teachers, traditions, institutions, and machines. This perspective aligns with sociological analyses of AI in education that emphasize the redistribution of agency and the reconfiguration of pedagogical power [34, 41]. By embedding disclosure practices into classroom use, the standard supports forms of reflexive pedagogy in which interpretive responsibility remains visible and contestable, ensuring that AI-assisted analysis supplements rather than supplants recognized sources of scholarly and religious authority.

6.3 Implications for platforms: disclosure as design norm

At the platform level, MRS-DST challenges designers and service providers to treat disclosure as a core design principle rather than a peripheral policy statement. Search engines, conversational interfaces, and educational applications increasingly operate as institutional amplifiers of authority, shaping what users perceive as canonical, credible, or normatively acceptable. Research on social transparency and algorithmic accountability shows that trust in AI systems emerges from explanations embedded within organizational and cultural contexts rather than from technical interpretability alone [12, 14]. By translating governance domains into interface-level markers, such as corpus provenance indicators, model limitation notices, and human validation signals, platforms can recalibrate user reliance and resist the drift toward algorithmic subjectivity, where machine-generated outputs are tacitly treated as authoritative [38].

6.4 Implications for policy: culturally grounded AI governance in Asia

For policymakers, particularly within Asian religious and educational contexts, MRS-DST offers a model of culturally embedded AI governance that moves beyond universalist ethical principles. Rather than applying abstract norms of fairness or transparency in isolation, the framework operationalizes accountability through locally meaningful epistemic practices, including *tafsīr, syarḥ, sanad–matn* reasoning, and traditions of scholarly validation. This orientation reflects broader arguments that governance must be grounded in culturally situated epistemic frameworks rather than treated as universally transferable design principles [19]. By positioning disclosure regimes, institutional review, and public accountability as primary levers of oversight, MRS-DST aligns with governance approaches that integrate regulatory, technical, and sociocultural dimensions of AI stewardship within distributed responsibility ecosystems [13].

6.5 Theoretical contribution: society shaping the algorithm

Research in Digital Religion has demonstrated that digital technologies reshape religious authority by transforming

communicative practices and patterns of religious participation [8]. Subsequent work has further shown how technological mediation reconfigures relationships between religious institutions, digital platforms, and faith communities [9]. Building on these foundations, this article explains how interpretive authority is institutionally constituted when AI-mediated systems participate in the production, validation, and circulation of religious knowledge. It synthesizes [37] theory of legitimate authority, Pitkin's [30] account of authorized representation, Saward's [32] theory of representative claims, Jasanoff's [18] conception of civic epistemology, Daston and Galison's [11] analysis of objectivity as an institutional practice, Floridi et al.'s [13] principles of AI ethics, and Novelli et al.'s [26] governance framework for responsible AI to develop an integrated theory of AI-mediated interpretive authority. MRS-DST operationalizes this theoretical synthesis through an institutional governance framework that supports accountable AI-assisted religious interpretation.

These implications underscore the central claim of this study: AI systems are shaped as much by institutions, traditions, and governance arrangements as they shape society. Although developed for AI-mediated religious interpretation, this principle extends to other domains in which AI contributes to the production of socially consequential knowledge, including law, education, medicine, journalism, and public policy. MRS-DST formalizes this reciprocal relationship by embedding cultural epistemologies and institutional accountability throughout the lifecycle of AI-mediated interpretation, from corpus construction and model design to publication, pedagogy, and platform deployment. In doing so, the framework responds to growing concerns about "algorithmic truth" and the redistribution of epistemic authority in digital knowledge systems [36]. Rather than resisting the growing role of AI in religious and cultural domains, MRS-DST situates technological capability within established moral, scholarly, and institutional ecologies, ensuring that AI functions as *musāʿid* (an assistive interpreter) rather than *muḥakkim* (an authoritative adjudicator).

6.6 Practical contribution

Beyond its theoretical contribution, this study proposes a practical governance framework for AI-assisted religious interpretation. MRS-DST translates the concept of interpretive authority into a structured reporting architecture that enables journals, educational institutions, digital platforms, and research organizations to document textual provenance, methodological traceability, algorithmic mediation, institutional validation, and ethical oversight. Rather than functioning as an ethics checklist, the framework operationalizes institutional accountability by making interpretive authority transparent, reviewable, and contestable throughout the lifecycle of AI-mediated knowledge production.

6.7 Field contribution

Although developed for AI-mediated religious interpretation, the concept of interpretive authority has broader relevance wherever AI participates in producing socially consequential knowledge. The framework therefore contributes not only to Digital Religion but also to wider debates in AI governance, Science and Technology Studies, and responsible AI. By distinguishing computational capability from institutional authority, the study offers a conceptual vocabulary for examining legitimacy in law, medicine, education, journalism, and public policy, where AI increasingly mediates interpretation without assuming responsibility for its consequences.

7 Limitations and scope

This study is intentionally situated within Islamic interpretive traditions, drawing primarily on *tafsīr*, *sharḥ*, and *sanad–matn* frameworks as institutional models of interpretive authority. Although the MRS-DST framework is designed to be transferable across culturally grounded knowledge systems, it does not claim universal applicability. Its conceptual architecture reflects the normative commitments embedded in Islamic scholarly practice. Accordingly, its application to other religious or cultural traditions should be regarded as an empirical question requiring contextual and institutional validation rather than assumed through theoretical generalization. This position is consistent with Asad's [4] argument that epistemic authority is constituted within historically specific discursive traditions. It also accords with Jasanoff's [19] conception of civic epistemologies, which emphasizes that practices of knowledge validation are culturally embedded and institutionally organized.

A second limitation arises from the rapid evolution of large language models, platforms, and deployment environments. Changes in training regimes, alignment strategies, interface design, and platform governance can alter the institutional conditions under which interpretive authority is enacted and disclosed. As a result, the procedural standards proposed here should be understood as adaptive rather than fixed, requiring periodic revision in response to technological and organizational shifts that reshape the interpretive landscape.

Finally, this study does not undertake systematic cross-tradition comparison. Religious and philosophical traditions such as Buddhism, Confucianism, Hinduism, and

Christianity possess distinct institutional arrangements for validating interpretation, including monastic lineages, canonical councils, ecclesiastical hierarchies, and pedagogical orders. Future research should examine how the MRS-DST domains translate into these alternative governance ecologies, and whether comparable disclosure regimes can be developed that respect their specific epistemic and institutional logics [4].

8 Conclusion

This study has argued that AI systems participate in the reconfiguration of interpretive authority rather than functioning solely as neutral instruments of scriptural retrieval or analysis. They shape who is authorized to speak, what counts as legitimate interpretation, and how religious knowledge circulates across institutions, platforms, and publics. As sacred texts are transformed into datasets, models, and interfaces, authority is no longer located exclusively within scholars, traditions, or communities. Instead, it is institutionally distributed across sociotechnical arrangements that embed assumptions about canon, method, and credibility into the infrastructures of interpretation.

Against this backdrop, governance cannot be reduced to engineering interventions such as debiasing, optimization, or explainability at the level of model performance alone. While necessary, such measures remain insufficient to address the institutional and cultural stakes of digital hermeneutics. What is at issue is not only how accurately a system represents a text, but how disclosure regimes, review structures, and cultural values determine the conditions under which interpretations are produced, validated, and authorized. Accountability therefore resides in the public visibility of interpretive pathways: who curated the corpus, how models delimit legibility, where human judgment intervenes, and which institutions assume responsibility for endorsement and sanction.

MRS-DST is proposed as a response to this governance challenge. Conceived as a living standard rather than a static compliance framework, it is designed to evolve with disciplinary norms, technological developments, and cultural epistemologies. By formalizing disclosure across textual, methodological, algorithmic, and ethical domains, the framework anchors AI-mediated interpretation within institutional ecologies of responsibility, ensuring that machines remain assistive participants in knowledge production rather than unexamined arbiters of meaning.

The broader implication is a call for accountable digital hermeneutics as a shared institutional project. As AI systems increasingly mediate access to sacred and culturally normative knowledge, the central question is no longer whether these systems can be rendered objective or fair in technical terms alone. It is whether societies can design and sustain the institutional conditions under which technological power is made transparent, contestable, and culturally grounded. The adoption of MRS-DST across journals, classrooms, platforms, and policy environments offers a pathway toward ensuring that the future of AI-mediated interpretation remains ethically governed, socially situated, and responsive to the diverse traditions of meaning-making it now inhabits.

Appendix A

Supporting documentation for the qur'anic stylistics case (illustrative MRS-DST audit record)

This appendix provides the supporting documentation referenced in Box 1. It records the corpus provenance, edition metadata, segmentation protocol, and exegetical resources underpinning the illustrative application of MRS-DST to AI-assisted Qur'anic stylistic analysis (Tables 3, 4, 5, and 6).

Table 3 Corpus specification sheet

Item	Specification
a. Corpus	The Holy Qur'an
b. Reading tradition	*Ḥafṣ ʿan ʿĀṣim*
c. Digital edition	Tanzil Uthmani text
d. Repository	Tanzil project
e. Language	Classical Arabic

Table 4 Edition metadata

Metadata element	Specification
a. Orthographic standard	Uthmani orthography
b. Verse numbering	Tanzil standard
c. Variant reading policy	Computational analysis restricted to the *Ḥafṣ ʿan ʿĀṣim* reading
d. Treatment of other *Qirāʾāt*	Alternative *qirāʾāt* were documented but excluded from computational clustering

Table 5 Segmentation protocol

Processing stage	Procedure
a. Unit of analysis	Individual Qur'anic verses
b. Segmentation boundary	Official Tanzil verse numbering
c. Tokenisation method	Sentence Piece sub-word segmentation
d. Interpretive responsibility	Computational procedures were limited to descriptive stylistic processing. All stylistic and theological interpretations remained the responsibility of human researchers

Appendix B

Supporting documentation for the *Hadīth* variant stability case (illustrative MRS-DST audit record)

This appendix provides the supporting documentation referenced in Box 2. It records the corpus provenance, edition metadata, source selection rationale, and interpretive resources underpinning the illustrative application of MRS-DST to AI-assisted *ḥadīth* variant stability analysis (Tables 7, 8, 9, and 10).

Table 6 Citation log of *Tafsīr* sources

Exegetical source	Purpose in the study
a. *al-Ṭabarī, Jāmiʿ al-Bayān*	Primary authority for validating lexical and contextual interpretation
b. *Ibn ʿĀshūr, al-Taḥrīr wa al-Tanwīr*	Primary authority for validating rhetorical and stylistic interpretation

Table 7 Corpus provenance sheet

Item	Specification
a. Primary collections	*Ṣaḥīḥ al-Bukhārī, Ṣaḥīḥ Muslim*, and selected Sunan collections
b. Primary editions	*Dār Ṭawq al-Najāh (Ṣaḥīḥ al-Bukhārī); Dār Iḥyāʾ al-Turāth (Ṣaḥīḥ Muslim)*
c. Editorial tradition	Critical published editions recognised in contemporary *ḥadīth* scholarship
d. Language	Classical Arabic

Table 8 Edition metadata

Metadata element	Specification
a. Editorial Basis	Critical published editions used throughout the study
b. Variant Inclusion Policy	Only *isnād*-linked textual variants documented in recognised classical sources were included
c. Treatment of Non-canonical Variants	Non-canonical digital variants were excluded from computational clustering

Table 9 Source justification memo

Documentation item	Justification
a. Inclusion criteria	Variants linked to recognised *isnād* traditions and documented in authoritative classical sources
b. Exclusion criteria	Unverified online texts and non-canonical digital variants were excluded
c. Source selection rationale	Selected collections represent the principal reference works in Sunni *ḥadīth* scholarship and provide the editorial basis for the illustrative analysis

Table 10 Classical sources used for interpretive validation

Reference	Purpose in the study
a. *Ṣaḥīḥ al-Bukhārī*	Primary corpus for transmission comparison
b. *ṢaḥīḥMuslim*	Comparative corpus for transmission comparison
c. *Ibn Ḥajar, Fatḥ al-Bārī*	Primary authority for validating transmission variants and editorial context
d. *al-Nawawī, Sharḥ Ṣaḥīḥ Muslim*	Primary authority for validating textual variants and interpretive conclusions

Acknowledgements The authors used a large language model as a writing support tool during the preparation of this manuscript to assist with language refinement, structural organization, and the drafting of preliminary text. All substantive conceptual development, theoretical framing, interpretation, and final editorial decisions were carried out by the authors. The authors reviewed, verified, and are solely responsible for the accuracy, originality, and integrity of the content.

Author contributions Hishomudin wrote the main manuscript text and Norfarhana prepared all the supplies material such as table and figure. All authors reviewed the manuscript.

Funding Open access funding provided by The Ministry of Higher Education Malaysia and Universiti Sains Islam Malaysia

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Al-*Shāṭibī*, I.: *Al-Muwāfaqāt fī uṣūl al-sharīʿah. Dār Ibn ʿAffān* (2004)
2. Amoore, L.: Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. Duke University Press (2020)
3. Anoraga, B. (2025). The future of artificial intelligence in/and of Islam: A view from Muslim Southeast Asia. Journal of Islamic and Muslim Societies, 9(1). https://doi.org/10.2979/jims.00031
4. Asad, T.: The idea of an anthropology of Islam. Georgetown University, Occasional Papers Series (1986)

5. Atallah, F.A.: Digital mediation and fatwa authority in contemporary Islam: a critical Islamic legal and media-theoretical framework. Religions 17(3), 350 (2026)
6. Bowker, G.C., Star, S.L.: Sorting Things Out: Classification and Its Consequences. MIT Press (1999)
7. Brown, J.A.C.: Hadith: Muhammad's Legacy in the Medieval and Modern World. Oneworld (2009)
8. Campbell, H.A.: Digital Religion: Understanding Religious Practice in New Media Worlds. Routledge (2013)
9. Cheong, P.H.: Authority. In: Campbell, H.A. (ed.) Digital Religion: Understanding Religious Practice in New Media Worlds, pp. 72–87. Routledge (2013)
10. Cooper, G., Tang, K.-S., Rappa, N.A.: Generative artificial intelligence as epistemic authority? Perspectives from higher education. In: Crompton, H., Burke, D. (eds.) Artificial Intelligence Applications in Higher Education: Theories, Ethics, and Case Studies for Universities, pp. 106–122. Routledge (2024). https://doi.org/10.4324/9781003440178-7
11. Daston, L., Galison, P.: Objectivity. Zone Books (2007)
12. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M., & Weisz, J. (2021). Expanding explainability: Towards social transparency in AI systems. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411764.3445188
13. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Vayena, E.: AI4People—an ethical framework for a good AI society. Minds Mach. 28(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
14. Gillespie, T.: The relevance of algorithms. In: Gillespie, T., Boczkowski, P., Foot, K. (eds.) Media Technologies: Essays on Communication, Materiality, and Society, pp. 167–194. MIT Press (2014)
15. Hallaq, W.B.: An Introduction to Islamic Law. Cambridge University Press (2009)
16. Haraway, D.: Situated knowledges: the science question in feminism and the privilege of partial perspective. Feminist Stud. 14(3), 575–599 (1988)
17. Hila, A.: The epistemological consequences of large language models: rethinking collective intelligence and institutional knowledge. AI Soc. 40(1), 1–15 (2025)
18. Jasanoff, S.: States of Knowledge: The Co-Production of Science and Social Order. Routledge (2004)
19. Jasanoff, S.: The Ethics of Invention: Technology and the Human Future. W. W. Norton (2016)
20. Latour, B.: Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press (1999)
21. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press (2005)
22. Longino, H.: Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press (1990)
23. Mahmood, S.: Politics of Piety: The Islamic Revival and the Feminist Subject. Princeton University Press (2005)
24. Malik, S.A.: Artificial intelligence and Islamic thought: two distinctive challenges. J. Islamic Muslim Stud. 8(2), 89–112 (2023)
25. Milella, F., Cabitza, F.: Perceiving AI as an epistemic authority or algority: a user study on the human attribution of authority to AI. Mach. Learn. Knowl. Extr. 8(2), 36 (2026). https://doi.org/10.3390/make8020036
26. Novelli, C., Taddeo, M., Floridi, L.: Accountability in artificial intelligence: what it is and how it works. AI Soc. 39(4), 1871–1882 (2023). https://doi.org/10.1007/s00146-023-01635-y
27. Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. Journal of Strategic Information Systems, 34(1), 101885. https://doi.org/10.1016/j.jsis.2024.101885
28. Papakostas, C. (2025). Artificial intelligence in religious education: Ethical, pedagogical, and theological perspectives. Religions, 16(5), 563. https://doi.org/10.3390/rel16050563
29. Pickering, A.: The Mangle of Practice: Time, Agency, and Science. University of Chicago Press (1995)
30. Pitkin, H.F.: The Concept of Representation. University of California Press (1967)
31. Reinhart, A.K.: Before Revelation: The Boundaries of Muslim Moral Thought. SUNY Press (2010)
32. Saward, M.: The representative claim. Contemp. Polit. Theory 5(3), 297–318 (2006). https://doi.org/10.1057/palgrave.cpt.9300234
33. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of FAT '19* (pp. 59–68). ACM. https://doi.org/10.1145/3287560.3287598
34. Selwyn, N.: Should Robots Replace Teachers? AI and the Future of Education. Polity Press (2019)
35. Serres, M. (2007). The parasite (L. R. Schehr, Trans.). University of Minnesota Press. (Original work published 1980)
36. Shin, D.: Automating epistemology: how AI reconfigures truth, authority, and verification. AI Soc. 41, 1553–1559 (2026). https://doi.org/10.1007/s00146-025-02560-y
37. Star, S.L., Griesemer, J.R.: Institutional ecology, "translations" and boundary objects. Soc. Stud. Sci. 19(3), 387–420 (1989)
38. Sun, Y., Xu, C.: Hybrid cognitive authority and algorithmic subjectivity: rethinking knowledge management in AI-driven communication. J. Knowl. Manag. (2025). https://doi.org/10.1108/JKM-02-2025-0252
39. Wahid, S.H., Abdulloh, S.: Digital Islamic authority: a revolutionary shift or mere hype? A systematic review. SN Soc. Sci. 6(2), 62 (2026). https://doi.org/10.1007/s43545-026-01328-5
40. Weber, M. (1978). Economy and society: an outline of interpretive sociology (G. Roth & C. Wittich, Eds.). University of California Press
41. Williamson, B.: Big data in education: the digital future of learning, policy and practice. SAGE Publ. (2017). https://doi.org/10.4135/9781529714920

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.