

Istiqlal-e-Marifat: An Iqbalian Framework for Epistemic Justice in Artificial Intelligence

Ammar Younas, Ph.D.

Director, Center for Cultural Governance, Beijing Institute of AI Safety and Governance,
Beijing, China

doctorammaryounas@mailsucas.ac.cn

Muhammad Ahsan Sikander

COMSATS University Islamabad, Islamabad, Pakistan

sp24-bds-002@isbstudent.comsats.edu.pk

Abstract

When AI systems are trained on datasets drawn predominantly from Western, English-speaking, secular-rationalist corpora, they do not merely encode information, they absorb one civilization's answers to the most fundamental questions of moral life: what constitutes a good life, what obligations bind human beings to one another, and how human experience is narrated and valued. We argue that this process constitutes an epistemic displacement rather than a mere problem of data representation. To address this problem philosophically, we introduce "*Istiqlal-e-Marifat*", sovereignty of deep knowing, as a novel philosophical concept derived from the thought of Allama Muhammad Iqbal. Using a conceptual and interpretive methodology, we reconstruct this principle from three dimensions of Iqbal's framework: the dynamism of *Khudi* at the civilizational level, his critique of stagnant and borrowed thought, and his affirmation of Eastern and Islamic intellectual traditions as bearers of irreducible epistemic resources. We apply this framework to three structural mechanisms of coded morality in AI, reinforcement learning from human feedback, tokenization reduction, and model collapse, showing how each enacts precisely the form of epistemic displacement Iqbal diagnosed. We find that *Istiqlal-e-Marifat* yields three normative implications: AI training must preserve epistemic structures rather than merely diversify data, governance must position communities as active epistemic partners rather than data sources, and evaluation must adopt moral benchmarks drawn from plural traditions rather than dominant world-models alone.

Keywords: Allama Iqbal, *Istiqlal-e-Marifat*, *Khudi*, AI ethics, moral epistemology, coded morality, cross-cultural AI

In the last decade, artificial intelligence has moved out of the research lab and into the lives of billions of people. AI helps doctors diagnose diseases, assists judges in sentencing decisions, determines who is granted an interview for an office job, decides which news articles we read; and

mediates innumerable exchanges between ordinary citizens and their governments. For most users, these systems simply operate as tools, appearing as neutral calculating machines that accept input and produce output. In this paper, we argue that the appearance of neutrality is philosophically flawed; the rectification of this philosophical error requires the help of an older thinker who wrestled with precisely this issue (in a different context) a century ago.

In their practice, these systems do not simply store information; in trained large language models, they absorb the assumptions, values, and narratives that have been woven into human-generated text. They learn how particular civilizations answer the essential questions of living: What makes for a good life? What responsibilities do we owe the strangers that we meet on the street? When can we justly harm one person to help many? What is fairness and to whom must it apply? These questions do not have universal, neutral answers; civilized people, different religious traditions and traditions of philosophical inquiry have come to radically different understandings that go beyond variations on a theme; these are distinct moral universes (Henrich et al., 2010). So, when we train large language models on a dominant source of this text, largely English, Western, and secular-rationalist, we do not train for universal intelligibility; we train for one civilized understanding.

This condition is known technically as data saturation. Researchers have observed that the rate of improvement on large language models decreases once models have had too much of their training data (Younas & Zeng, 2026b). When this occurs, the standard response is simply to search for new sources of data. But the paper argues that this is a mistake, since the saturation at play is not the saturation of tokens or of semantic patterns, but the saturation of moral imagination. This occurs when an AI model's ingrained assumptions about human living and being have been fixed into a particular civilization's model so deeply that the capacity to conceive of alternative moralities is annihilated. The scale of this issue has been confirmed empirically: models trained on dominant corpora systematically undervalue the moral commitments of non-Western people, even when prompted in their own languages (Atari et al., 2024). This demonstrates that AI is not learning human morality, but one sliver of humanity's many understandings and that sliver is mistakenly interpreted by the system as a whole.

Current work has explored various facets of this problem through algorithmic bias, decolonial AI, and indigenous data sovereignty (Birhane, 2021; Mohamed et al., 2020; Tuhiwai

Smith, 1999), but none has undertaken to provide a systematically philosophical account of a civilization's ability to be in command of its own living morality, nor the loss of this ability either through historical dependency or encoded imitation. In this paper, we argue that Allama Muhammad Iqbal does exactly this, offering an *Iqbalian* philosophy that can provide an account precise and original enough to make a novel contribution to both Iqbal studies and the philosophy of AI.

This problem also has a broader genealogy in philosophical critiques of knowledge and modernity. Critical theorists have long examined how dominant forms of rationality can reduce diverse modes of understanding into narrower technical or instrumental frameworks (Horkheimer & Adorno, 1947/2002; Marcuse, 1964). Similar concerns appear across Eastern and non-Western intellectual traditions: Confucian approaches emphasize relational accounts of human flourishing, Buddhist philosophies challenge purely individualistic conceptions of agency, and Indigenous epistemologies emphasize situated relationships between knowledge, community, and lived experience. However, these approaches have generally remained within debates concerning inclusion, representation, or alternative ethical perspectives. What distinguishes Iqbal's contribution is that he provides a systematic philosophical account of a civilization's capacity to remain the author of its own living moral knowledge.

The principle of *Istiqlal-e-Marifat*, "sovereignty of deep knowing," we will derive from *The Secrets of the Self (Asrar-e-Khudi)*, *The Reconstruction of Religious Thought in Islam*, and *The Message of the East (Payam-e-Mashriq)*. Bringing together the Arabic terms *Istiqlal* ("independence," "sovereignty," "self-determination") and *Marifat* ("deep," "situational," "transformative moral knowing," as distinct from simple information or *Ilm*), it offers a principle that asserts a civilization's living tradition of intellectual and moral knowledge is an irreducible source of its ability to know morality, and that its erosion—whether through colonial subjugation or the imitative encoding of AI, represents an irresolvable philosophical crisis.. We will develop the concept from Iqbal's principal works, apply it to the ways that moral knowing can become encoded into the logic of AI, and conclude with some proposals for the regulation and design of artificial intelligence systems.

Research Objectives

This paper pursues three objectives. First, it constructs *Istiqlal-e-Marifat* as a novel philosophical concept by reconstructing the principle from three foundational texts of Iqbal's philosophy: *Asrar-e-Khudi*, *The Reconstruction of Religious Thought in Islam*, and *Payam-e-Mashriq*. Second, it applies this concept to the problem of coded morality in AI, showing how the technical mechanisms of large language model training — data saturation, RLHF, tokenization, and model collapse — enact the form of epistemic displacement Iqbal diagnosed in a different historical context. Third, it derives normative implications from the *Istiqlal-e-Marifat* framework for the training, governance, and evaluation of AI systems, with the aim of contributing a philosophically grounded non-Western framework to the field of cross-cultural AI ethics.

This paper does not attempt to construct an “Iqbalian morality” for artificial intelligence or translate Iqbal’s complete ethical philosophy into computational principles. Rather, it extracts one epistemological principle from Iqbal’s thought, the sovereignty of living moral knowledge and applies it as a framework for evaluating whether AI systems preserve or displace diverse traditions of understanding.

Methodology

This paper employs a conceptual and interpretive philosophical methodology. The first stage is textual reconstruction: primary sources from Iqbal's philosophy are read closely and systematically to identify the philosophical components of what this paper terms *Istiqlal-e-Marifat*. This reconstruction proceeds immanently — the concept is derived from within Iqbal's own framework rather than imposed from external categories — and draws on the full body of his major philosophical works rather than selectively citing passages in isolation.

The second stage is conceptual application. The reconstructed framework is brought into contact with the technical and structural literature on large language model development, AI ethics, and cross-cultural bias in AI systems. This stage is analytical rather than empirical: it identifies structural homologies between Iqbal's diagnosis of epistemic dependency and the mechanisms of coded morality in AI, and does not make or test quantitative claims.

These parallels are philosophical rather than historical equivalences. The paper does not claim that Iqbal anticipated contemporary artificial intelligence; rather, it argues that his analysis

of intellectual dependency provides conceptual resources for understanding similar structures of epistemic displacement in contemporary AI systems.

The third stage is normative derivation. Following the method of applied philosophy, the paper draws practical implications for AI training, governance, and evaluation from the *Istiqlal-e-Marifat* framework. These implications are assessed against existing literature in AI ethics and epistemic justice rather than against empirical data.

This methodology is appropriate to the paper's aims. The contribution is philosophical, the introduction of a novel concept and its application to a contemporary problem and does not depend on empirical findings that would require a different methodological apparatus.

Istiqlal-e-Marifat: Constructing the Concept from Iqbal's Philosophy

The concept of *Istiqlal-e-Marifat*, although not an actual phrase of Iqbal's, represents a theme and a principle that he elaborates throughout his philosophical endeavor. *Istiqlal* bears the full and proper weight of sovereignty in Arabic, Urdu and Persian discourse: "independence, autonomy, self-governance". *Marifat* denotes profound, concrete, experienced knowing within the tradition of Islamic philosophy and Sufism:

The knowledge which becomes itself a condition that transforms the subject into himself; a knowledge which is formed through immersion in a tradition and its origins and which cannot be reduced to a body of propositions or the mere passing on of information.

Istiqlal-e-Marifat represents this sovereignty of experience and knowledge within one's own tradition: the condition in which a civilization will constitute its own moral and existential understanding within its own life, rather than simply draw from ready-made answers provided by an external authority on knowledge.

This distinction reflects a central feature of Iqbal's philosophical project: knowledge is not merely the accumulation of inherited propositions but an active process through which the self develops creative engagement with reality. Intellectual dependency therefore emerges when a tradition ceases to generate interpretations through its own intellectual resources and merely reproduces categories produced elsewhere.

The concept of *Istiqlal-e-Marifat* has three dimensions which constitute it, each with grounding in Iqbal's key texts:

The first dimension is the vitality of Khudi at the civilizational level. In *Asrar-e-Khudi*, Iqbal establishes that the self's moral force derives from active engagement with its own resources, not from passive reception of inherited formulas or external frameworks:

When life gathers strength from the Self,
The river of life expands into an ocean.
(Iqbal, 1915/2001, p. 23; Nicholson translation)

The individual Khudi is defined as selfhood and constitution through active intellectual and moral work. For Iqbal, collective Khudi is “self-creation” in the civilizational sense. A civilization that possesses its collective Khudi produces its own responses to the central moral dilemmas of human existence. A civilization that has lost it either repeats transmitted dogmas devoid of life and understanding or accepts the solutions of a stronger foreign civilization as its own. The first state of the civilization is that it possesses vitality, the third state that it adopts the solutions of a foreign civilization as its own in addition to ceasing the generation of independent thought is the deepest form of intellectual dependency that *Istiqlal-e-Marifat* diagnoses.

Importantly, Iqbal's concept of Khudi does not imply isolation from other intellectual traditions. His own philosophical development involved sustained engagement with Western thinkers, including Bergson, Whitehead, Nietzsche, and Goethe. The issue is therefore not intellectual exchange itself, but passive imitation that prevents creative reconstruction and independent moral reflection.

The second sense is Iqbal's focus on the generation of its own internal movement as the condition of the civilization's health. "Practically Stationary" Iqbal characterizes the situation of Muslim thought in *The Reconstruction* (Iqbal, 1930/2013, p. 7), not a state of ignorance but of following paths carved in the past without forging one's own by drawing from one's internal vitality. His characteristic warning is: "Our only fear is that the dazzling exterior of European culture may arrest our movement and we may fail to reach the true inwardness of that culture." (Iqbal, 1930/2013, p. 7)

The epistemic significance of Iqbal's warning should not be overlooked, and its philosophical rigor is substantial. He is not against engaging with European civilization; he has engaged with it seriously; he has, in fact, studied it thoroughly in Cambridge and Munich; he has, among others, taken inspiration from Bergson, Whitehead, Nietzsche, and Goethe. His warning is only that one be not 'stopped' by its surface; he is afraid one will be stopped by the symptomatic outcomes of a living intellectual tradition rather than the tradition itself and lose both its original locomotion and the potential for truly internalizing what the foreign tradition has to offer. *Istiqal-e-Marifat* is precisely the name for not being 'stopped.' It means the condition by virtue of which an individual or civilization can engage with the other without losing its own intellectual movement.

The third dimension of this particular insistence of Iqbal is a consistent commitment to the epistemic traditions of East and the Islamic world as dynamic resources of *Marifat* not substitutable by Western rationality. In *Payam-e-Mashriq*, written as a philosophical response to Goethe's West-Östlicher Divan and as a declaration of the East's own epistemic standing — Iqbal writes:

"The Westerner's knowledge has no cognizance of Self,
Eastern wisdom has illumined the world."

(Iqbal, 1923/1994, *Payam-e-Mashriq*; translator's rendering)

This isn't cultural nationalism; rather, it is a metaphysical claim regarding the non-reducible nature of traditions of *Marifat*. The epistemic tradition of the East, based on Tawhid, Sufi ontology of self and ultimate being, Islamic jurisprudence of autonomous reasoning, and Eastern philosophical traditions in general, possesses moral resources concerning human existence that cannot be fully captured by Western rationalist categories alone. *Istiqal-e-Marifat* refers to the principle of keeping these traditions as living, active resources, and refusing their absorption into external systems. The contemporary relevance of this aspect of Iqbal's thought has been noted in recent scholarship. The persistence of Muslim and Eastern traditions of knowledge, and the presence of their unique philosophical resources in relation to present-day dilemmas (Hussain, 2025; Rodas, 2023), have been highlighted in recent scholarship.

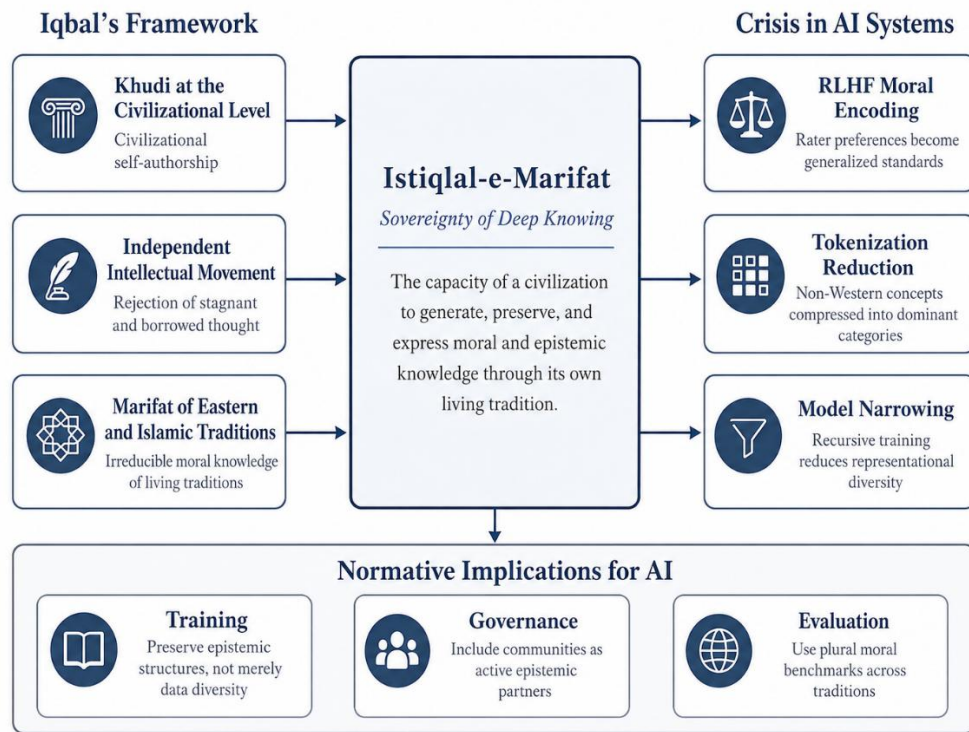
Iqbal's diagnosis is inseparable from a positive account of moral development. For Iqbal, the good is not conformity to a fixed external code but the creative realization of Khudi through active moral engagement. His discussion of *ijtihad* in *The Reconstruction* presents moral knowing as a dynamic process of interpretation and reconstruction rather than passive repetition of inherited formulations (Iqbal, 1930/2013; Diagne, 2010). It is this movement-centred conception of moral knowledge, the capacity to generate responses to changing conditions while remaining rooted in a living tradition that *Istiqlal-e-Marifat* brings into the contemporary AI context.

These three elements form the unique philosophical concept that is *Istiqlal-e-Marifat*. It is clearly distinguished from contemporary notions of epistemic sovereignty, which focus on the issues of rights to data, governance and political self-determination (Tuhiwai Smith, 1999; Carroll et al., 2020), because it operates at the level of living moral epistemology. The form of sovereignty at stake here is the capacity to produce moral and existential knowledge. Sovereignty here doesn't pertain to the right to control one's own data; instead, it pertains to moral epistemology, which is central to the problem of coded morality.

This distinguishes *Istiqlal-e-Marifat* from existing accounts of epistemic injustice. Fricker's account focuses primarily on individuals being wronged in their capacity as knowers within existing epistemic structures, particularly through credibility deficits or exclusion from knowledge exchange (Fricker, 2007). *Istiqlal-e-Marifat* operates at a deeper and collective level: it concerns whether communities possess the capacity to generate their own moral categories of understanding. The central concern is therefore not only misrecognition within a shared epistemic framework, but displacement of the generative sources through which moral knowledge is produced.

Figure 1

Istiqlal-e-Marifat: An Iqbalian Framework for AI Ethics



Note. Author's conceptualization based on Iqbal (1915/2001; 1923/1994; 1930/2013).

Coded Morality and the Saturation of Human Narrative: The Technical Mechanisms

Coded morality refers to a result of data saturation: the encoding of a specific civilization's assumptions about morality into the operating system of AI. To grasp how this comes about, it is necessary to provide a brief, intuitive account of how most contemporary AI systems are designed, assuming most readers of this journal do not work as AI engineers.

Table 1Technical Mechanisms of Coded Morality and Their *Iqbalian* Dimensions

Mechanism	Technical Process	Epistemic Effect	<i>Iqbalian</i> Dimension
Training Data Bias	Dominant corpora shape model representations of human values and norms	Particular moral frameworks become universalized	Loss of epistemic autonomy (<i>Khudi</i>)
RLHF Moral Encoding	Human preferences calibrate outputs according to situated normative assumptions	Cultural judgments become embedded as general standards	Substitution of inherited categories for active reconstruction
Tokenization Reduction	Culture-specific concepts (<i>Adl</i> , <i>Ghairat</i> , <i>Ren</i>) are mapped onto dominant linguistic structures	Moral meanings are compressed or distorted	Reduction of <i>Marifat</i> into transferable information
Model Narrowing	Recursive training may reduce representational diversity	Alternative moral perspectives become progressively marginalized	Arrest of intellectual movement

Note. Conceptual synthesis based on Bender et al. (2021), Atari et al. (2024), Shumailov et al. (2024), and Younas and Zeng (2025a, 2026b).

Large language models, the underlying AI systems powering things like ChatGPT, Gemini, and others, work by training on vast corpuses of human text. The system learns statistical relationships between words and phrases; what moral conclusions tend to co-occur with what other conclusions in what contexts; which ways of describing human experiences are statistically common. The system does not "understand" this data in the way a human reader does, but learns to generate outputs statistically similar to that data. But this training data is not value-neutral. It represents the specific assumptions, values, and modes of describing human existence dominant in the corpus from which it was drawn; and such corpuses are overwhelmingly Western, English, and secular-rationalist. What results is a system that does not encode universal human moral truths,

but the most statistically prominent moral values present in its training data (Bender et al., 2021; Gabriel, 2020).

A second layer of training, reinforcement learning from human feedback (RLHF), further complicates matters. Human raters evaluate system outputs, their preferences are then used to adjust the system's reward function (i.e., to program the system to respond in certain ways, which are deemed good by human raters), and thus the human moral preferences of the raters become the criterion for a good output. RLHF processes rely on human evaluators whose judgments are shaped by particular linguistic, cultural, and institutional contexts. In many prominent systems, evaluation has historically relied heavily on English-speaking and Western-based annotation environments. Their specific moral intuitions become the system's benchmark of quality for output that is used globally by individuals with entirely disparate moral traditions. Again, what is encoded is not a universal ethical framework, but a culturally specific moral intuition applied globally as neutral judgement.

Recent alignment approaches, including Constitutional AI, attempt to reduce reliance on direct human preference labeling by guiding models through explicit principles rather than conventional reinforcement learning from human feedback alone (Bai et al., 2022). However, this does not eliminate the underlying epistemic question: the principles selected as constitutional guidance still reflect particular assumptions about harm, fairness, responsibility, and acceptable behaviour.

A third layer of operation occurs at the linguistic level itself. Language parsers, the systems that process text for the machine, are trained primarily on English data, meaning non-Western languages are processed inefficiently; in more extreme cases, non-Western concepts with no close parallel in the dominant training language are simply non-existent in the model, or collapsed onto rough English approximations that misrepresent them. The Arabic *Adl*, or justice as cosmic equilibrium; the Urdu *Ghairat*, or honor connected with communal identity; and the Confucian *Ren*, or benevolence as relationally contingent virtue, are not simply translated. They are mapped onto an existing schema not built to receive them (Younas & Zeng, 2025a). The effect is not that AI extends its reasoning into new areas of moral inquiry, but that it reduces those areas to its existing schema. This is what Iqbal foresaw: intellectual exploration arrested by sheer display of a foreign tradition, its own lexicon substituted for one borrowed from another culture.

Substantial evidence of these structural effects already exists. Large language models reliably default toward a Western moral frame even when prompted in non-Western languages (Atari et al., 2024). Studies applying the Moral Foundations Questionnaire to nineteen cultural groups have demonstrated that LLMs fail to represent these different moral worlds, regardless of their linguistic abilities (Münker, 2025). The original critical paper on LLMs confirmed that they reflect the cultural biases of their training distributions, often in non-recoverable ways (Bender et al., 2021). Work published in PNAS shows that models demonstrate exaggerated biases on moral reasoning tests; that they not only reflect the biases of their training data but multiply them (Cheung et al., 2025). This exacerbation of existing biases is the technological analogue of Iqbal's account of the stunting of the mind and the slide into dependence: the system not only reproduces the epistemic weaknesses of its origin but amplifies them with each cycle.

The process continues with model collapse, the term for the phenomenon of systems trained on recursive output from other systems gradually converging on more constrained outputs (Shumailov et al., 2024). When AI systems produce training data for subsequent systems, the cultural biases contained within the original corpus do not just persist; they become intensified as the heterogeneity of the training data shrinks. In Iqbal's terms, this can be read as a form of civilizational narrowing accelerated by technological iteration: a process that, by reinforcing a specific moral universe, crowds out other such universes. This process not only leaves the *Marifat* of others unrepresented; it erases it altogether, to be replaced by an intensifying, purified version of one tradition's moral worldview.

This argument should not be understood as a claim that all forms of synthetic-data training inevitably produce technical model collapse (Shumailov et al., 2024; Gerstgrasser et al., 2024). The concern here is epistemic narrowing: even where systems continue functioning technically, repeated reliance on limited representational patterns may reduce the diversity of moral perspectives available within AI systems. From an Iqbalian perspective, the central danger is therefore not technological failure but the gradual arrest of intellectual movement.

At the policy level, this dynamic plays out explicitly within the major AI ethical frameworks themselves. The adoption of five shared principles (transparency, justice, non-maleficence, responsibility, and privacy) across multiple domains seems to suggest consensus (Jobin et al., 2019), but this reflects the prevalence of Western liberal political theory in the

discussions surrounding these guidelines, rather than any truly cross-cultural convergence of values. Different traditions express these five principles in the context of entirely different moral universes, understand them as based on entirely different accounts of the human subject, and weigh them against each other according to entirely different priorities (Coeckelbergh, 2023; Wong, 2020; Zhu, 2022). The veneer of shared principles, masking stark differences in moral substance, is what the data saturation process effaces, and what *Istiqal-e-Marifat* refers to as the usurpation of living moral traditions.

***Istiqal-e-Marifat* as a Normative Framework for AI**

Istiqal-e-Marifat is therefore not just a diagnosis but a normative direction. The way Iqbal reacted to epistemic dependence was through reconstruction: active, self-possessed reconstruction of lived traditions in response to contemporary needs. Three consequences follow from this framework for AI development:

The first consequence is about the training of AI. *Istiqal-e-Marifat* says that inclusion of diverse traditions into the AI training process is not about representation alone, but about epistemic integrity. Texts from such traditions may be included in training data without attention to their epistemic structure: their vocabularies, styles of argumentation, and moral framings. In that case, they risk being reified and assimilated into the categories of the dominant paradigm. Their *Marifat* will only be reduced to information. The argument of Iqbal that the civilizations of the Islamic world and Asia possess inherent and non-reducible moral wisdom, translates to saying that real epistemic diversity in AI requires those traditions to become active epistemic contributors in training data, not additional data points (Younas & Zeng, 2026a).

Previous works argue that cultural corpora should be understood as containers of alternative epistemic world-models and not as extra data points (Younas & Zeng, 2026b), and *Istiqal-e-Marifat* gives us the philosophical justification as to why this distinction matters and the risks of ignoring it.

The second consequence concerns the governance of AI. Intellectual reconstruction as conceived by Iqbal was essentially a collective practice: the retrieval of *Istiqal-e-Marifat* in Muslim and Eastern thought required active thinking and writing on the part of the thinkers engaged in these traditions, and not reception of external frames of thought. This point directly

translates to AI governance: the communities whose epistemic traditions are engaged by AI should become active partners in its design and evaluation, not simply sources of extracted data or passive recipients of symbolic representation (Younas & Zeng, 2024).

The Moral Grounding Model posits an equivalent principle at the education level: moral formation must precede and constitute technical skill (Younas & Zeng, 2025b); the application to AI development implies that epistemic grounding should precede the construction of the AI architecture and be derived from active engagement with traditions in consideration, not from external models. This ties in nicely with the emerging work on epistemic justice in AI, where a major demand is that the epistemic traditions of non-Western world-models are taken seriously as possessing legitimate epistemic authority (Birhane, 2021; Mohamed et al., 2020).

The third consequence is about evaluation. Iqbal's own intellectual activity as depicted in *The Reconstruction* offers a template for truly intercultural epistemic engagement: it draws from Western thought, the intellectual traditions of the Islamic world, and modern science without diminishing any in the face of the others. Applied to AI evaluation, this means that performance metrics cannot be solely drawn from the traditions of the dominant paradigm; an AI system should not be considered morally competent merely because it meets the standards set by dominant world-models. The standard set by *Istiqlal-e-Marifat* demands that a truly competent AI moral agent can engage with the moral demands of various traditions, and this requires more than the ability to translate their propositions into familiar terms, but an ability to incorporate and respond to their respective *Marifat*.

Conclusion

In this paper, we introduced *Istiqlal-e-Marifat*, sovereignty of deep knowing, as a novel philosophical concept from Iqbal's long meditation on the conditions under which civilizations maintain or lose their ability to independently produce autonomous, self-authored moral and existential thought. Building upon the concept of *Khudi* at the civilizational level, his diagnosis of intellectual dormancy in *The Reconstruction*, and his argument in *Payam-e-Mashriq* of Asia and the Islamic world as carriers of irreducible *Marifat*, we offered a concept that names, with philosophical accuracy, what is at stake when AI systems ingest dominant corpora and push out the moral and existential discourses of other civilizations.

This paper has shown that these technical mechanisms (RLHF encoding of culturally-specific rater preferences, tokenization which reduces non-Western moral ideas to the predominant linguistic categories, and the tendency for model collapse to amplify original cultural biases through successive training stages) are not merely failure modes of implementation. They are structural features of an epistemic structure not designed for *Istiqlal-e-Marifat*. These systems do not merely risk diluting the *Marifat* of Eastern, Islamic, and other non-Western traditions; they may progressively displace it, reinforcing the statistical regularities of dominant moral frameworks while presenting them as universal intelligence.

This paper makes three distinct contributions. First, it introduces *Istiqlal-e-Marifat* as a concept derived from Iqbal's primary texts and applies it to a contemporary problem he could not have foreseen, explaining epistemic autonomy not merely as political independence but as the capacity of traditions to generate their own moral and existential categories. Second, it presents Iqbal's thought as a generative resource for AI ethics. Third, it argues that recovering *Istiqlal-e-Marifat* in AI development requires serious engagement with living traditions as sources of moral knowledge. We present Iqbal's thought as a serious and generative resource for the domain of AI ethics, not as a source of Islamic content to be tacked onto a pre-existing framework, but as a thinker whose notion of civilizational epistemic health offers both vocabulary and vision. We then argue that an effort to recuperate *Istiqlal-e-Marifat* in the context of AI development must involve the sort of serious intellectual work Iqbal himself modeled: critical engagement with living traditions as sources of active, critical moral knowledge rather than merely content. His warning that the seductive exterior of an overpowering culture has the potential to stop us in our tracks remains potent, and in an era of artificial intelligence, it carries a weight, scale, and audience Iqbal would have fully understood.

Future research may extend this framework in several directions. One possible avenue is the development of practical evaluation models for AI systems based on the principle of *Istiqlal-e-Marifat*, especially in relation to culturally distinct moral concepts and forms of reasoning. Further studies may also examine how non-Western epistemic traditions are represented, reduced, or displaced in multilingual AI systems. In addition, future work can explore governance models in which communities and scholars rooted in different intellectual traditions participate directly in the design, training, and assessment of artificial intelligence.

References

- Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2024). Which humans? *PNAS Nexus*, 3(7). <https://doi.org/10.1093/pnasnexus/pgae258>
- Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, M., Ndousse, K., van der Poel, A., Jose, A., & Amodei, D. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv. <https://doi.org/10.48550/arXiv.2212.08073>
- 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* (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), 100205. <https://doi.org/10.1016/j.patter.2021.100205>
- Carroll, S. R., Garba, I., Figueroa-Rodriguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, 19(1), 43. <https://doi.org/10.5334/dsj-2020-043>
- Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25), e2412015122. <https://doi.org/10.1073/pnas.2412015122>
- Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: Political epistemology in times of artificial intelligence. *AI and Ethics*, 3, 1341–1350. <https://doi.org/10.1007/s43681-022-00239-4>
- Diagne, S. B. (2010). *Islam and open society: Fidelity and movement in the philosophy of Muhammad Iqbal*. CODESRIA. <https://www.jstor.org/stable/j.ctvk3gnz2>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198237907.001.0001>
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., & Koyejo, S. (2024). *Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data*. arXiv. <https://doi.org/10.48550/arXiv.2404.01413>

- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.
<https://doi.org/10.1017/S0140525X0999152X>
- Horkheimer, M., & Adorno, T. W. (2002). *Dialectic of enlightenment: Philosophical fragments* (E. Jephcott, Trans.). Stanford University Press. (Original work published 1947)
- Hussain, R. (2025). A critical re-exploration of Dr. Muhammad Iqbal's philosophical, political, and economic contributions. *Islamic Political Thought*, 12(2), 129–151.
<https://doi.org/10.30497/ipt.2025.77869>
- Iqbal, M. (1915/2001). *Asrar-e-Khudi* [The secrets of the self] (R. A. Nicholson, Trans.). Sh. Muhammad Ashraf. (Original work published 1915)
- Iqbal, M. (1923/1994). *Payam-e-Mashriq* [Message of the East]. Iqbal Academy Pakistan. (Original work published 1923)
- Iqbal, M. (1930/2013). *The reconstruction of religious thought in Islam* (J. Majeed, Ed.). Stanford University Press. (Original work published 1930)
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Marcuse, H. (1964). *One-dimensional man: Studies in the ideology of advanced industrial society*. Beacon Press.
- Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684. <https://doi.org/10.1007/s13347-020-00405-8>
- Münker, S. (2025). Cultural bias in large language models: Evaluating AI agents through moral questionnaires. In *Proceedings of the Symposium on Moral and Legal AI Alignment (IACAP/AISB Conference 2025)*. <https://arxiv.org/abs/2507.10073>
- Rodas, C. (2023). The challenge of Muhammad Iqbal's philosophy of Khudi to Ibn Arabi's metaphysical anthropology. *Religions*, 14(5), 683. <https://doi.org/10.3390/rel14050683>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Tuhiwai Smith, L. (1999). *Decolonizing methodologies: Research and Indigenous peoples*. Zed Books.
- Wong, P.-H. (2020). Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. *Philosophy & Technology*, 33(4), 705–715.
<https://doi.org/10.1007/s13347-020-00408-5>

- Younas, A., & Zeng, Y. (2024). A philosophical inquiry into AI-inclusive epistemology. In *Proceedings of the 8th Asian Conference on Artificial Intelligence Technology (ACAIT 2024)* (pp. 177–181). IEEE. <https://doi.org/10.1109/ACAIT63902.2024.11022320>
- Younas, A., & Zeng, Y. (2025a). *Ghairat: Navigating cultural morality in patient care*. *Journal of Medical Ethics*. Advance online publication. <https://doi.org/10.1136/jme-2025-111188>
- Younas, A., & Zeng, Y. (2025b). Moral grounding before algorithms: A cross-cultural critique of AI education in schools. *AI & Society*. <https://doi.org/10.1007/s00146-025-02749-1>
- Younas, A., & Zeng, Y. (2026a). Towards culture driven artificial intelligence to bridge the cultural cognition gap. *Discover Artificial Intelligence*, 6, 252. <https://doi.org/10.1007/s44163-026-01060-2>
- Younas, A., & Zeng, Y. (2026b). Beyond data saturation in large language models: Cultural corpora and the expansion of epistemic world-models. *Journal of Cultural Cognitive Science*. <https://doi.org/10.1007/s41809-026-00204-1>
- Zhu, J. (2022). AI ethics with Chinese characteristics? Concerns and preferred solutions in Chinese academia. *AI & Society*, 39(3), 1127–1140. <https://doi.org/10.1007/s00146-022-01578-w>