



Conscience after Code: AI, Moral Authority, and the Limits of External Governance

Muhammad Faiz ul Rehman, Muhammad Khalid

Idara-tul-Mustafa International, Gujranwala

METADATA

Paper history

Received: 29 March 2026

Revised: 02 April 2026

Accepted: 29 April 2026

Published online: 01 June 2026

Corresponding author

Email: mfaizjaami@gmail.com

(Muhammad Faiz ul Rehman)

Keywords

AI ethics

Moral authority

Taqwa

Amanah

Maqasid al-shariah

Islamic ethics

Citation

Rehman MF and Khalid M (2026)

Conscience after code: AI, moral authority, and the limits of External Governance. *Al-Amīn (Faiṣalābād)* 4: 26040101.

<https://doi.org/10.63793/Al-Amin/31>.

ABSTRACT

This article examines a central limitation in contemporary AI ethics: the growing sophistication of external governance has not been matched by an equally serious account of internal moral authority. Current frameworks have advanced important principles of safety, transparency, accountability, fairness, and human oversight, yet they remain focused primarily on regulating observable conduct. They do not fully address how morally responsible judgement is formed in the human agent who designs, deploys, governs, and uses technological systems. The article argues that this gap becomes especially significant in the age of generative and increasingly agentic AI, where external systems may assist, influence, and sometimes displace human deliberation. In response, the study develops an Islamic ethical framework grounded in Quranic anthropology and the moral concepts of taqwa, amanah, ihsan, niyyah, muraqabah, muhasabah, and maqasid al-shariah. Its central claim is that external code cannot replace internal conscience. AI may contribute to efficiency, coordination, and public benefit, but it cannot inherit moral authority. The article therefore proposes that a just and sustainable AI order requires a clear hierarchy: technology must remain an assistant to conscience-led human agency, while the ultimate grounding of safety, responsibility, and restraint lies in the integrated moral vision of the Islamic tradition.

INTRODUCTION

Artificial intelligence has become one of the most consequential forces of the contemporary age, influencing governance, education, media, research, finance, and everyday human conduct. Current ethical and regulatory frameworks rightly stress safety, transparency, fairness, accountability, and human oversight. Yet these frameworks remain largely concerned with the external regulation of action. They do not fully address the prior question of how morally responsible judgement is formed within the human being who designs, authorizes, and uses such systems.

This article argues that the deeper crisis of the AI age is not technology itself, but the displacement of moral

authority from conscience to code. From the Qur'anic perspective, true moral order is not secured by surveillance, procedure, or outward compliance alone, but by inward accountability before Allah. The Qur'an (2183) states:

يا أيها الذين آمنوا كتب عليكم الصيام كما كتب على الذين من قبلكم لعلكم تتقون

"O you who believe, fasting is prescribed for you, as it was prescribed for those before you, so that you may attain Taqwa"

Likewise,

إن أكرمكم عند الله أتقاكم

"Indeed, the most noble of you in the sight of Allah is the most God-conscious of you" (Quran 49:13).

The Prophet صلى الله عليه وسلم said:

إنما الأعمال بالنيات

“Actions are only by intentions,”

And

أن تعبد الله كأنك تراه فإن لم تكن تراه فإنه يراك

“Worship Allah as though you see Him; and if you do not see Him, He sees you.”

AI may assist humanity, but it cannot replace morally formed conscience under divine guidance.

The literature relevant to this study may be arranged around four converging discussions: first, modern analyses of surveillance, external regulation, and the procedural management of conduct; second, recent AI ethics literature on fairness, accountability, transparency, and human oversight; third, wider philosophical concerns about the preservation of human judgement and moral agency in the age of intelligent systems; and fourth, Islamic sources and contemporary Muslim scholarship on *taqwa*, *amanah*, *ihsan*, and moral self-accountability. The central contention emerging from this review is that contemporary scholarship has become highly developed in diagnosing external risk and designing regulatory response, but remains comparatively weaker in accounting for the inward formation of moral authority (MacIntyre, 2007). A major strand of modern critical literature shows how late-modern governance increasingly shifts from the cultivation of character to the administration of observable behavior. In Michel Foucault’s analysis, disciplinary systems do not merely punish; they organize visibility, normalize conduct, and produce governable subjects. Although his analysis predates digital AI, its relevance has only deepened in algorithmic environments where ranking, monitoring, and behavioral prediction are embedded into everyday infrastructures. This literature is important because it helps explain how ethical life may be increasingly externalized into systems of observation, measurement, and correction.

This concern is intensified in Shoshana Zuboff’s account of surveillance capitalism, where human experience itself becomes raw material for data extraction, prediction, and behavioral modification. Her analysis is crucial for the present study because it shows that digital systems do not merely record action after the fact; they seek to shape future action by converting conduct into a manageable object of commercial and institutional power. Yet even in this highly influential literature, the problem is framed largely in terms of rights, autonomy, privacy, and market domination, not in terms of the deeper moral anthropology required to restore inward accountability (Zuboff, 2019).

A parallel philosophical diagnosis appears in MacIntyre’s critique of modern moral discourse. His central argument is that modernity retains fragments of ethical language while losing the teleological and communal traditions that once made virtue intelligible. This insight is highly relevant to AI ethics, because

contemporary governance often presupposes that procedural safeguards can compensate for the erosion of shared moral formation. MacIntyre helps clarify why systems may become more efficient in regulating conduct while societies become less capable of forming persons whose judgements are ordered by virtue (MacIntyre, 2007).

Within AI ethics proper, the dominant literature has focused on fairness, accountability, transparency, explainability, privacy, and human oversight. This body of work has achieved major conceptual and institutional importance by identifying bias, opacity, discrimination, safety failures, and concentration of power as central problems of AI deployment. Such work remains indispensable, especially because it translates ethical concern into operational governance. At the same time, much of this literature remains concerned with the design of better constraints, better auditing, and better oversight mechanisms, rather than with the formation of morally serious human agents who must ultimately interpret, govern, and bear responsibility for technological systems (Floridi and Cows, 2019).

This regulatory orientation is evident in influential institutional frameworks as well. The UNESCO Recommendation on the Ethics of Artificial Intelligence and the NIST AI Risk Management Framework both foreground values such as human rights, accountability, transparency, robustness, and human oversight. These frameworks are of immense practical value, and the present study does not oppose them. Rather, it observes that they primarily operate at the level of institutional design and risk governance. They tell us how systems should be constrained, evaluated, and supervised, but they do not fully answer how the supervising human subject is to be morally formed (UNESCO, 2021).

A newer line of scholarship has begun to move closer to the present article’s concern by focusing on the preservation of human judgement under conditions of AI assistance. Here the question is no longer limited to fairness or transparency, but extends to whether habitual reliance on AI may erode interpretation, deliberation, and agency themselves. This concern is especially important because the moral problem of AI is not exhausted by harmful outputs; it also includes the possibility that human beings become progressively less capable of serious judgement by delegating interpretive and evaluative labor to machines (Kim, 2025).

Recent Muslim scholarship has also begun to address AI from within Islamic categories, though the field remains emergent. Some discussions focus on permissibility, harm, and public interest; others attempt to build broader ethical models through *amanah*, *maqasid al-shariah*, and *khalifah*. What makes this developing literature especially valuable is that it does not treat the human being merely as a rights-bearing individual or a node within regulatory systems, but as a morally

responsible servant whose knowledge, action, and intention remain accountable before Allah. This gives Islamic ethics a conceptual advantage when the question shifts from external compliance to internal authority.

Among the most important published discussions for the present study is Mohammad Azram's treatment of governance in Islam. His argument is significant because it shows that Islamic governance cannot be reduced to formal rule or institutional procedure; rather, it is tied to the formation of a psycho-moral attitude of submission to Allah, service to moral order, and accountability before divine judgement. This is directly relevant to the current article, since it supports the claim that governance in the Islamic horizon is never purely external. Its ultimate ground lies in the disciplined moral interior of the human being (Azram, 2012).

This inward dimension is even clearer in the classical ethical tradition. Al-Muhasibi, al-Ghazali, and Ibn al-Qayyim all treat the heart, intention, self-accounting, and vigilance over the nafs as indispensable to moral life. In this tradition, ethical action is not secured merely by visible conformity, but by muhasabah, muraqabah, ikhlas, and taqwa. The self is not presumed reliable merely because it acts within rules; it must be trained, examined, corrected, and oriented toward Allah. This tradition therefore offers a rich grammar for thinking about moral formation in a way that exceeds the limits of procedural compliance (al-Ghazali, 1993).

The Quranic category that gathers these strands most powerfully is taqwa. In Quranic usage, taqwa is not merely fear in a narrow sense, nor merely legal caution; it is a sustained condition of God-consciousness, restraint, vigilance, and answerability. It is this concept that gives Islamic ethics a particularly strong response to the problem identified in the modern literature. Where surveillance-based systems depend upon external visibility, taqwa remains operative in concealment as well as publicity. Where algorithmic governance may optimize behavior, taqwa concerns the interior relation between act, intention, and divine accountability (Izutsu, 2002). The literature review therefore yields a clear conclusion. Modern critical theory and AI governance scholarship provide powerful diagnoses of surveillance, institutional control, and technological risk, but they do not sufficiently resolve the problem of inward moral formation. Islamic ethical thought, by contrast, supplies a much thicker anthropology of responsibility through taqwa, ihsan, amanah, muraqabah, and muhasabah. The gap to be addressed by the present article lies precisely here: how to bring contemporary AI ethics into serious conversation with a Quranically grounded account of internal moral authority, without collapsing either into technophobia or into mere proceduralism (Hashas, 2015).

MATERIALS AND METHODS

This article adopts a normative interdisciplinary methodology. It is normative because its central concern is not only how AI systems function, but how they ought to be governed in relation to human responsibility, justice, and moral authority. It is interdisciplinary because the problem under discussion cannot be adequately addressed from one field alone. AI governance literature identifies real concerns such as opacity, safety, overreliance, and the erosion of meaningful human judgement, yet the deeper question of inward moral formation requires engagement with moral philosophy, Quranic anthropology, Hadith, and the classical Islamic tradition of ethical self-discipline. In this sense, the article does not treat Islamic sources as decorative supplements to technical debate; rather, it brings them into direct analytical conversation with contemporary concerns about agency, accountability, and the preservation of judgement (Kim, 2025).

Methodologically, the article proceeds through conceptual analysis, close textual reading, and comparative ethical interpretation. Conceptual analysis is used to distinguish between external governance and internal moral authority, between procedural compliance and inward accountability, and between regulation of conduct and formation of conscience. Close textual reading is used in relation to the Qur'an, Hadith, and selected primary Islamic ethical works in order to reconstruct the moral logic of taqwa, ihsan, muraqabah, and muhasabah. Comparative ethical interpretation is then employed to place these Islamic categories into dialogue with current AI ethics discourse without forcing either tradition into the language of the other. The aim is therefore neither apologetic harmonization nor civilizational denunciation, but a disciplined inquiry into whether present AI ethics can remain adequate without a stronger account of the human interior (Izutsu, 2002).

The primary textual base of the article consists of the Qur'an, canonical Hadith, and selected classical works of tafsir and ethics. Qur'anic verses are treated as sources of moral anthropology, not merely as pious citations. The Hadith corpus is used especially where it defines inward accountability with conceptual precision, as in the Hadith of Jibril and other narrations relating to intention, ihsan, and hidden obedience. Classical tafsir is consulted where the ethical architecture of taqwa, fasting, justice, and trust can be clarified through authoritative exegetical tradition. In addition, selected ethical works by scholars such as al-Ghazali, al-Muhasibi, and Ibn al-Qayyim are drawn upon where they illuminate the practices of vigilance, self-accounting, disciplining the nafs, and purification of intention. This allows the article to ground its claims in primary Islamic materials rather than in generalized moral rhetoric (al-Ghazali, 1993).

Alongside these primary Islamic sources, the article engages selectively with modern scholarly and policy literature on AI. The operative word here is selectively. The article does not attempt a comprehensive survey of every technical branch of AI research, nor does it attempt to provide legal commentary on all emerging regulatory regimes. Rather, it focuses on those strands of recent literature that bear directly on its central concern: surveillance, behavioral management, external governance, human oversight, deskilling, and the preservation of meaningful judgement. The purpose of this engagement is to identify the strengths of contemporary AI ethics while also testing its limits. This is important because the article does not deny the value of regulation, risk management, or human oversight; it argues only that these remain incomplete unless they are joined to an account of inward moral responsibility (UNESCO, 2021).

The scope of the article is therefore deliberately focused. It is not a technical paper on model architecture, not a legal brief on AI regulation, not a sociological account of platform capitalism in its entirety, and not a general treatise on Islamic governance. Its concern is more specific: the moral limits of external governance in the age of AI, and the relevance of *taqwa* as a model of internal moral authority. For that reason, the discussion of present AI interventions remains limited to what is necessary for the argument, namely those forms of AI that expand automation, recommendation, decision-support, and the external organization of conduct. Likewise, the discussion of Islamic thought remains limited to those concepts that directly illuminate moral formation, accountability, and the relation between outward order and inward discipline. This limitation is not a weakness but a condition of analytical clarity (Azram, 2012).

The article also adopts an important argumentative restraint. It does not claim that religious ethics alone can replace institutional safeguards, nor does it claim that AI is intrinsically destructive. It does not reject oversight, legal regulation, or technical governance. Rather, it argues that these belong to the outer architecture of order and that they become morally fragile when detached from inwardly formed agents. In this respect, the article's use of Islamic sources is constructive rather than reactionary. *Taqwa* is presented not as an anti-technological slogan, but as a disciplined account of inner accountability that may supplement present discussions of governance by addressing what external systems cannot produce for themselves (IBID). Finally, the methodological horizon of the article is comparative but not relativistic. It recognizes that secular virtue ethics, human-rights discourse, and institutional governance each provide real and serious resources for AI ethics. Yet it also maintains that

Quranic anthropology offers a thicker conception of accountability because it binds action, intention, concealment, and answerability before Allah into one moral structure. The article's wager is that this structure can illuminate a major omission in current debates: the tendency to assume that if systems are sufficiently audited, supervised, and regulated, the ethical problem has been substantially solved. The present study challenges that assumption by arguing that no adequate ethics of AI can ignore the formation of the one who judges, chooses, restrains, and answers (Kim, 2025).

THE PRESENT AI MOMENT: INTERVENTIONS, BENEFITS, AND RISKS

The present AI moment is marked not only by rapid technical advance but by a widening expansion of where and how AI systems intervene in human life. What began as narrower forms of automation and prediction has moved into generative, multimodal, and increasingly agent-like systems capable of producing text, images, code, summaries, recommendations, and coordinated task execution across domains. OECD's recent conceptual study on agentic AI notes that current discussions increasingly distinguish between conventional AI systems and more autonomous, goal-directed agents that can plan, use tools, and act across sequences of tasks with reduced human prompting. This shift matters because the moral question is no longer confined to isolated outputs; it increasingly concerns delegated processes, distributed action, and the management of human judgement within expanding technical environments (OECD, 2026).

These interventions are now visible across governance, education, health, research, finance, media, and everyday communication. Public institutions use AI for triage, fraud detection, administrative classification, and service delivery. Educational settings use it for tutoring, translation, summarization, and personalized support. Workplaces use it for drafting, coding assistance, scheduling, search, and productivity enhancement. Health-related environments increasingly rely on AI for imaging support, transcription, pattern recognition, and workflow assistance. UNESCO's ethical framework remains relevant precisely because it recognizes that AI is no longer a marginal tool but a general-purpose technological layer capable of affecting rights, dignity, access, and social organization at scale (UNESCO, 2021). It is important, therefore, to state clearly that the present article does not approach AI as an intrinsically harmful phenomenon. The benefits are real and substantial. Generative and multimodal systems can improve access to knowledge, support translation across languages, reduce routine burdens, assist research, widen accessibility for disabled users, accelerate drafting and synthesis, and increase institutional efficiency in contexts where time and resources are limited. In some domains,

AI can enhance rather than diminish human capacity by handling repetitive tasks and freeing attention for higher-order deliberation. NIST's Generative AI Profile reflects this balanced posture: it does not deny the utility of generative AI, but seeks to govern its use through structured risk management so that beneficial deployment is not separated from responsible practice (NIST, 2024).

At the same time, the expansion of AI interventions has deepened a corresponding field of risks. Some are now familiar: bias, opacity, hallucination, privacy loss, discrimination, data insecurity, and concentration of power. Others are more subtle but equally significant: overreliance on machine outputs, deskilling of human judgement, diffusion of responsibility, erosion of deliberative habits, and the gradual normalization of systems that organize conduct through prediction and recommendation rather than moral reasoning. NIST's profile repeatedly stresses that generative AI introduces distinctive risks tied to synthetic content, misinformation, unpredictability, and the difficulty of evaluating outputs at scale. These are not merely technical inconveniences; they affect the conditions under which truth, trust, and responsibility are publicly negotiated (*ibid*).

Recent governance developments show that regulators increasingly recognize both the promise and the danger of these systems. The European Union's AI Act has moved into phased application, and by August 2025 obligations for general-purpose AI models became effective, including requirements concerning documentation, copyright policy, and additional duties for models presenting systemic risk.

A further feature of the present AI moment is the emergence of systems described as agents or agentic AI. Here the concern is not simply that a model generates an answer, but that it may pursue goals through multiple steps, coordinate tools, maintain context across tasks, and act with limited immediate oversight. OECD's 2026 paper is especially useful because it does not treat "agentic AI" as a slogan; rather, it identifies recurring features in the literature such as autonomy, goal-directedness, planning, memory, and tool use. The more these features become operational in mainstream systems, the more urgent become questions about supervision, accountability, escalation, and the preservation of meaningful human judgement. What is at stake is not merely efficiency, but the moral location of agency within human-technical arrangements (OECD, 2026).

For that reason, the present AI moment should not be described in simplistic terms of either celebration or alarm. It is a mixed condition of genuine promise and genuine peril. AI can assist learning, widen access, improve coordination, and support human work. It can also intensify surveillance, weaken interpretive discipline, amplify irresponsible delegation, and normalize the illusion that better systems can compensate for thinner moral agency. UNESCO's ethical framework is

instructive here because it continues to insist on human dignity, proportionality, fairness, transparency, and human oversight, thereby resisting both technological utopianism and unreflective adoption. Yet the existence of such principles does not eliminate the deeper problem explored in this article. It only clarifies that the governance of AI, however necessary, remains incomplete if it does not ask what kind of human being is governing, deploying, trusting, and obeying these systems (UNESCO, 2021).

The significance of this section, then, is diagnostic. The present AI moment is defined by expanding interventions, real and often substantial benefits, and equally real risks that now exceed narrow technical categories. This mixed reality requires more than refusal and more than enthusiasm. It requires an ethical framework capable of affirming beneficial innovation while identifying the point at which external systems begin to displace rather than assist human moral responsibility. That is the threshold on which the next section turns: the limits of external governance itself (Floridi and Cowls, 2019).

EXTERNAL GOVERNANCE AND ITS LIMITS

The preceding discussion makes clear that the present article is not opposed to technological development, nor to the beneficial use of artificial intelligence in administration, education, research, health, and other domains of human need. The argument is more precise. Technical advancement is valuable when it remains ordered to human good, but this ordering depends upon a governing principle: technology must remain assistant, supporter, and instrument, not driver, ruler, or substitute for conscience. Once the relationship is reversed and external systems begin to lead while human moral judgement becomes reactive, dependent, or secondary, the ethical center of action is displaced. The danger, then, is not innovation as such, but the inversion through which instruments start to occupy the seat of moral direction (Heidegger, 1977). At the level of public ethics, external governance refers to those mechanisms through which conduct is supervised, constrained, audited, and standardized from outside the agent. In contemporary AI discourse, this includes regulation, compliance frameworks, technical safeguards, model evaluations, documentation duties, institutional oversight, and procedural controls. These mechanisms are not only legitimate but indispensable. Without them, technologically powerful systems can operate without accountability, magnify bias, intensify opacity, and expose persons and institutions to serious harm. For this reason, modern frameworks such as the NIST AI Risk Management Framework and UNESCO's ethical recommendation deserve to be read as necessary achievements in the governance of emerging technology (NIST, 2024).

Yet the necessity of external governance does not imply its sufficiency. This distinction is central. External mechanisms can regulate behavior, but they cannot independently generate the moral qualities required for responsible human action. They may deter abuse, but they cannot purify intention. They may increase traceability, but they cannot produce sincerity. They may require disclosure, but they cannot create reverence, humility, restraint, or moral seriousness. At most, they can establish conditions under which better conduct becomes more likely. They do not, by themselves, form the inner subject who must still judge, choose, interpret, refuse, and answer. That is why a purely compliance-centered moral order remains fragile: it depends on outer pressure while leaving the deeper springs of action underdeveloped (MacIntyre, 2007).

This limitation has long been visible in broader political and social theory. Foucault's analysis of disciplinary power shows that systems of surveillance, normalization, and examination can shape observable conduct with considerable efficiency. But the success of discipline in producing manageable behavior does not amount to the formation of virtuous persons. A subject may be compliant without being just, visible without being sincere, and controlled without being morally transformed. The relevance of this insight to AI governance is considerable. Where algorithmic systems increase the precision of monitoring and prediction, they may also intensify the illusion that the ethical problem is solved once conduct becomes externally legible and technically manageable (Foucault, 1995).

A similar insight may be drawn from contemporary analyses of digital power. In surveillance capitalism, as Zuboff argues, systems increasingly seek not only to know behavior but to shape it through prediction and modification. This produces a model of order in which the subject is increasingly acted upon by architectures of nudging, ranking, personalization, and anticipatory control. Such systems may become extremely effective in steering behavior, yet their effectiveness does not resolve the moral question of whether the governed subject has become wiser, more responsible, or more just. Indeed, the greater the success of external behavioral management, the greater the temptation to neglect the inner labor of character formation (Zuboff, 2019).

This problem becomes sharper in the age of AI because contemporary systems do not merely watch; they recommend, prioritize, summarize, infer, and increasingly coordinate action. With the rise of more agent-like systems, the temptation grows to confuse operational competence with moral adequacy. A system may outperform humans in speed, consistency, pattern recognition, and procedural recall, yet still remain incapable of bearing moral responsibility in the full sense. The issue is not simply that AI lacks human consciousness in a philosophical sense, but that moral

authority involves answerability, intention, accountability under concealment, and the possibility of self-restraint even where no immediate external force compels it. These are qualities of the moral subject, not merely of the procedure (Kim, 2025).

For this reason, the article argues that the most serious risk lies not only in flawed systems, but in the moral inversion through which human beings gradually surrender the driver's seat of judgement to external processes. When technology becomes leader rather than assistant, the relation between means and ends is destabilized. Tools are designed to serve purposes, but if the tool begins to define the horizon of decision, then purpose itself is slowly subordinated to optimization, convenience, or prediction. In such a condition, the human agent may continue to act, but increasingly as a functionary of systems whose operational logic is stronger than his own moral deliberation. The catastrophe feared here is not merely a dramatic collapse, but a gradual erosion of conscience under the prestige of technical capacity (Ellul, 1964).

It is therefore necessary to distinguish clearly between assistance and substitution. Assistance strengthens human action while leaving moral direction in human hands. Substitution occurs when the external system no longer supports judgement but increasingly displaces it. This distinction should govern the ethical use of AI in every field. A translation tool may assist understanding; it should not become a replacement for intellectual responsibility. A decision-support system may organize relevant information; it should not become the silent sovereign of human judgement. A recommendation engine may extend capacity; it should not become the hidden architect of values, priorities, and desires. In each case, the decisive question is whether human conscience remains the leader of action or has been reduced to the follower of a technical process (UNESCO, 2021).

The language of "human oversight" found in many current governance documents is therefore important but still too thin if not further specified. Oversight can mean little more than procedural review, formal approval, or post hoc supervision unless it is connected to a morally formed human agent. A human being who merely ratifies machine-guided decisions without seriousness of conscience does not solve the ethical problem simply by being present in the loop. Presence is not the same as responsibility. Human oversight becomes ethically meaningful only when the human overseer retains substantive judgement, moral independence, and the willingness to refuse technically efficient but morally unsound action. This is why the article insists that conscience must remain the binding principle governing every invention, intervention, revolution, and evolution. Without such priority, oversight becomes ceremonial and human leadership becomes nominal (Dignum, 2019).

At this point, the limits of external governance can be stated with precision. External systems can set boundaries, assign duties, improve accountability, reduce some harms, and create more responsible conditions of use. They cannot create the moral centre from which rightly ordered action proceeds. They cannot ensure that the one who designs, deploys, governs, or obeys the system possesses sincerity, restraint, humility, justice, or fear of answerability. For that reason, governance remains structurally incomplete unless joined to an account of inward authority. This is not an argument against regulation; it is an argument against the illusion that regulation alone can carry the entire burden of ethics. The next section therefore turns to the concept that addresses this omission most directly in the Islamic tradition: *taqwa* as internal moral authority (Azram, 2012).

QUR'ANIC ANTHROPOLOGY: TAQWA AS INTERNAL MORAL AUTHORITY

Any serious attempt to address the moral limits of external governance must eventually ask a prior question: what, in the Quranic view of the human being, makes responsible action possible even when surveillance is absent and compulsion is weak? The answer is not located merely in law, fear of punishment, or public reputation. It is located in the moral interior of the human being, in that cultivated state which the Qur'an repeatedly names as *taqwa*. For this reason, *taqwa* should not be reduced to a vague pious sentiment, nor narrowed to the language of caution alone. In Quranic anthropology, *taqwa* designates a disciplined condition of inward vigilance through which the human being becomes answerable before Allah in both public and concealed action. This inner answerability is what makes *taqwa* central to moral authority (Izutsu, 2002).

The linguistic foundation of *taqwa* is well known. Derived from the root w-q-y, it carries the sense of guarding, protecting, shielding, and placing a barrier between oneself and what is harmful or feared. Yet in Quranic usage the term undergoes ethical deepening. It no longer signifies mere avoidance in a physical or tactical sense; rather, it signifies moral self-protection through conscious obedience and restrained desire. Ibn Manzur (1997) explains *taqwa* in terms of making a *wiqayah*, a protective barrier, while al-Raghib al-Asfahani shows that in ethical usage it becomes the guarding of the soul against what leads to sin and destruction. The concept therefore already contains an anthropology: the human being is not morally secure by default, but must actively cultivate an inward shield against corruption, vanity, and disordered desire.

The Qur'an presents *taqwa* not as a secondary virtue but as an organizing principle of moral life. One of its most important formulations appears in the verse of fasting:

يَا أَيُّهَا الَّذِينَ آمَنُوا كُتِبَ عَلَيْكُمُ الصِّيَامُ كَمَا كُتِبَ عَلَى الَّذِينَ مِنْ قَبْلِكُمْ لَعَلَّكُمْ تَتَّقُونَ

“O you who believe, fasting is prescribed for you, as it was prescribed for those before you, so that you may become mindful of God”. (Haleem, 2004)

This verse is decisive because it links prescribed discipline to moral formation. The purpose of fasting is not stated as bodily denial for its own sake, nor as ritual conformity alone, but as the cultivation of *taqwa*. In other words, the Qur'an places one of the most visible acts of worship in the service of an invisible moral state. That structure is highly significant for the present study. It demonstrates that the divine pedagogy aims beyond external compliance toward the formation of an inward authority capable of governing the self when no worldly observer is present.

Al-Tabari's (2003) exegesis is especially important here because it preserves the ethical logic of the verse without reducing it to ceremonial observance. In his commentary on 2:183, he explains that fasting was prescribed so that believers might become among the *muttaqun* through restraint from what Allah has forbidden and through habituation in obedience. The exegetical move is crucial: *taqwa* is generated not by abstract instruction alone, but by disciplined acts that train the soul in unseen obedience. The result is a moral subject who does not require continuous external enforcement in order to refrain from wrong. This helps clarify why *taqwa* belongs to anthropology rather than merely to devotional vocabulary. It names a formed human condition (ibid).

Ibn Kathir (1999) develops the same point in a manner even more directly relevant to the problem of desire and moral restraint. In his commentary on the verse of fasting, he treats *saum* as a means of weakening the pathways of appetite and disciplining the impulses through which the *nafs* is often overwhelmed. *Taqwa* here emerges not as passive fear, but as active mastery under divine orientation. The human being becomes morally reliable not because appetite disappears, but because appetite is subordinated to obedience. This is of particular importance in an age of technological systems that often intensify immediacy, stimulation, and impulse satisfaction. The Quranic anthropology of *taqwa* does not deny human desire; it trains it under moral ends (Qur'an). The Qur'an also connects *taqwa* to human worth and moral distinction. In Surat al-Hujurat, the most cited verse on this theme declares:

إِنَّ أَكْرَمَكُمْ عِنْدَ اللَّهِ أَتْقَاكُمْ

“Indeed, the most noble of you in the sight of God is the most mindful of Him” (Haleem, 2004).

This verse carries two major anthropological implications. First, it denies that rank, power, lineage, ethnicity, or visible status are the final measures of human worth. Second, it relocates dignity into a moral and spiritual criterion that is inward before it is outward. The noblest human being is not the most technically equipped,

politically dominant, or materially empowered, but the one most deeply ordered by taqwa. In the context of the present article, this provides a powerful corrective to any civilizational framework that allows technological capacity to become the measure of value. The Qur'anic scale is different: quwwah without taqwa is unstable, whereas taqwa rightly orders power. The uploaded study on the taqwa-quwwah dichotomy is useful precisely because it shows that Muslim intellectual reflection has long wrestled with the danger of material strength detached from inner moral formation.

This distinction is not only social but existential. Taqwa transforms the human being from a creature managed from outside into a self-aware moral agent who carries an inward sense of answerability. That is why the Qur'an repeatedly joins taqwa to knowledge, remembrance, restraint, justice, and divine nearness. It is not merely a prohibition-focused attitude. It is a total moral orientation. Izutsu's semantic analysis is especially valuable on this point, because he demonstrates that Quranic ethical concepts do not function as isolated virtues but as interlocking components of a worldview in which moral consciousness is inseparable from ontological relation to Allah. Taqwa, therefore, is not simply one virtue among many; it is a governing disposition that shapes the moral reading of reality itself (Izutsu, 2002).

A further dimension appears in the verses that remind human beings of constant moral record and answerability:

إِذْ يَتْلَى الْمُتْلَفِينَ عَنِ الْيَمِينِ وَعَنِ الشِّمَالِ قَعِيدٌ . مَا يَلْفُظُ مِنْ قَوْلٍ إِلَّا لَدَيْهِ رَقِيبٌ عَتِيدٌ

“When the two receivers receive, seated on the right and on the left, not a word does one utter but there is a watcher ready beside him” (Haleem, 2004)

The significance of these verses lies not in producing paranoia, but in constructing moral seriousness. The human being is not abandoned to invisibility. Speech, intention, and deed unfold within a reality of witness. Yet the witness here is not a worldly regime of surveillance designed to dominate subjects; it is part of a divine moral order that calls the human being to self-accounting. The difference is profound. Modern surveillance operates externally and often instrumentally. Quranic accountability operates inwardly and ethically. The former may produce caution; the latter seeks conscience.

Muslim thinkers and philosophers have repeatedly emphasized that no just external order can be sustained if the inner order of the soul is neglected. Al-Muhasibi made self-accounting the center of moral reform because he understood that the crisis of action begins in the unexamined heart. Al-Ghazali similarly argued that the outward limbs follow the inward condition, and that the corruption of intention hollows out the apparent righteousness of deeds. Ibn al-Qayyim later deepened this logic by presenting muraqabah and muhasabah as conditions for sincerity, steadfastness, and moral

rectification. In all of these writers, the point is unmistakable: institutions matter, law matters, visible conduct matters, but the decisive battlefield remains the inner human being (al-Ghazali, 1993).

This same insight appears in modern Muslim scholarship on governance. Mohammad Azram's discussion of good governance in Islam is especially helpful because it defines taqwa as a psycho-moral attitude of accountability before Allah that disciplines passions under reason guided by revelation. He further argues that without such inward preparation the most formal structures of governance remain unstable, because the problem of corruption is ultimately rooted in the human subject who governs and is governed. This argument directly reinforces the thesis of the present article. If governance depends upon morally prepared persons, then no external regulatory architecture, however sophisticated, can bear the whole burden of ethics by itself. The inner authority of taqwa remains indispensable.

The philosophical force of taqwa becomes even clearer when placed beside the tendency of modern systems to treat visibility as if it were equivalent to accountability. The Qur'anic model resists that reduction. A person of taqwa remains bound by moral responsibility in khalwah as well as in jalwah, in concealment as well as in publicity. This is why fasting occupies such an important place in the Quranic and Prophetic moral imagination: it is obedience in the absence of visible enforcement. The one who fasts can break the fast in secrecy, yet refrains because taqwa has relocated authority inward. Here the anthropology of taqwa appears in its clearest practical form. The most decisive ethical restraint may occur precisely where external supervision is weakest.

The contrast with purely external governance is now sharp. External systems can monitor, deter, rank, and constrain. Taqwa forms a subject who watches over himself before being watched by others. External control organizes conduct through institutional force. Taqwa internalizes accountability through conscious relation to Allah. External frameworks are necessary for social order, but they remain dependent upon the moral quality of the agents who design, apply, and obey them. For that reason, the Quranic anthropology of taqwa does not compete with governance as such; rather, it identifies the deeper condition without which governance becomes shallow, selective, or performative. This is why taqwa should be understood here as internal moral authority. It is the inward government of the self that precedes and stabilizes the just government of systems (Hashas, 2015).

HADITH, AQWAL, AND ATHAR IN MORAL FORMATION

If the previous section established taqwa as the Quranic ground of internal moral authority, the Hadith corpus,

together with the *aqwal* and *athar* of the early Muslim tradition, shows how that authority is formed, refined, and sustained in lived moral practice. The Prophetic Sunnah does not treat moral excellence as a merely abstract state of belief, nor as a legal minimum secured by external compliance. Rather, it trains the believer to become answerable inwardly before Allah, vigilant over intention, restrained in desire, and serious in self-accounting. For that reason, Hadith is indispensable to the present argument, because it translates the Quranic anthropology of *taqwa* into an operative pedagogy of the heart, the tongue, the limbs, and the hidden self (al-Ghazali, 1993). The most concentrated Prophetic articulation of inward moral consciousness appears in the Hadith of Jibril:

ان تعبد الله كأنك تراه فان لم تكن تراه فإنه يراك

“That you worship Allah as though you see Him, and if you do not see Him, then indeed He sees you” (al-Ghazali, 1993).

This hadith is of decisive importance because it defines *ihsan* as a mode of inward awareness, not merely outward correctness. The first half expresses the highest moral and spiritual orientation, while the second grounds it in continuous divine surveillance that does not humiliate but ennoble the believer through answerability. Here moral formation reaches beyond public observance into conscious interiority. A human being shaped by *ihsan* does not require constant worldly monitoring in order to act rightly, because the deeper principle of supervision has already been internalized. This is why the hadith is central to any Islamic account of moral authority: it binds conduct to consciousness and consciousness to divine presence. The same logic appears in the foundational hadith on intention:

انما الاعمال بالنيات وانما لكل امرئ ما نوى

“Actions are only by intentions, and every person shall have only what he intended” (Sahih al-Bukhari).

This narration is methodologically vital for the present study because it establishes that the ethical value of action cannot be exhausted by outward form alone. Two acts may appear similar externally and yet differ fundamentally in moral worth because their inward orientation differs. In relation to AI ethics, this point is especially powerful. Systems may classify outputs, track processes, and evaluate compliance, but they cannot independently determine the sincerity, integrity, and moral seriousness with which a human agent acts. The hadith of intention therefore marks a limit of external governance: visible regularity does not automatically disclose inward truth. A further Prophetic principle clarifies the locus of corruption and reform:

ألا وان في الجسد مضغة إذا صلحت صلح الجسد كله وإذا فسد فسد الجسد كله ألا وهي القلب

“Truly, in the body there is a morsel of flesh: if it is sound, the whole body is sound; and if it is corrupted, the whole body is corrupted. Truly, it is the heart (Sahih al-Bukhari). The significance of this hadith is profound. It

places the moral center of human life not in external conformity alone but in the heart as the seat of orientation, motive, and corruption. Al-Ghazali made this hadith one of the structural foundations of his moral psychology, arguing that outward reform without inward rectification remains unstable and often deceptive. In the framework of the present article, this hadith directly supports the thesis that no amount of procedural governance can carry the whole ethical burden if the heart of the acting subject remains morally unformed (al-Ghazali, 1993).

The formative role of hidden obedience becomes especially clear in relation to fasting. In one hadith qudsi, Allah says:

كل عمل ابن آدم له الا الصوم فإنه لي وأنا اجزي به

“Every deed of the son of Adam is for him except fasting; it is for Me, and I alone shall reward it (al-Ghazali, 1993).

This hadith has long been understood by Muslim scholars as indicating a distinctive relation between fasting and sincerity. Because fasting is uniquely resistant to public verification in its inward reality, it becomes a school of concealed fidelity. Ibn Rajab explains that fasting is among the clearest acts through which *ikhlas* is tested, since the fasting person can secretly violate it yet refrains for the sake of Allah alone. This is analytically crucial for the present article. It shows that one of the most important modes of Islamic moral training is built precisely on the absence of worldly surveillance. In such a case, the governing force is not external compulsion but inward accountability (Sahih al-Bukhari). The Prophetic Sunnah also repeatedly links moral worth to restraint of speech and self-governance, not only to visible achievement. The Messenger of Allah صلى الله عليه وسلم said:

من كان يؤمن بالله واليوم الآخر فليقل خيرا أو ليصمت

“Whoever believes in Allah and the Last Day, let him speak good or remain silent (Sahih al-Bukhari).

This hadith is significant because it frames speech not as a spontaneous right detached from ethics, but as an arena of disciplined accountability. In contemporary technological environments where communication is accelerated, amplified, and often detached from reflection, this Prophetic standard retains particular force. The morally formed person is not the one who merely can speak, publish, or generate output, but the one who governs expression through responsibility and restraint. In that sense, the Sunnah does not merely condemn harmful speech; it cultivates a character capable of withholding it (Sahih al-Bukhari). The *aqwal* and *athar* of the early Muslim tradition deepen this inner pedagogy. Umar ibn al-Khattab is widely reported to have said:

حاسبوا انفسكم قبل ان تحاسبوا

“Take account of yourselves before you are taken to account”

Whether cited as *athar* or *hikmah*, the importance of this statement in Muslim ethical thought is unmistakable. It condenses the logic of *muhasabah* into one imperative:

the morally serious person does not wait for external exposure before beginning reform. He anticipates judgement through self-examination. Al-Muhasibi built an entire ethical program on this principle, insisting that the *nafs* must be questioned, corrected, and monitored so that action may be reordered before corruption hardens into habit. This early Islamic emphasis on self-audit is highly relevant to modern discussions of accountability, because it presents an interior form of review prior to institutional review (Al-Muhasibi, 1964).

The same moral grammar appears in the sayings of Hasan al-Basri, who described the believer as one who remains a guardian over himself, taking himself to account for Allah. The force of such statements lies in their anthropological realism. They assume that the human self is vulnerable to self-deception, appetite, heedlessness, and moral drift, and that ethical life therefore requires more than formal rule-following. It requires vigilance. In the language of the present article, they point to an important truth: systems may detect violations after they occur, but only inward discipline can prevent the normalization of a heart that no longer resists them (Al-Makki, 1997).

Muslim thinkers and philosophers extended this insight beyond devotional life into broader ethical theory. Miskawayh, for example, treated moral refinement as the disciplined habituation of the soul under the guidance of reason and virtue, while al-Ghazali joined spiritual psychology to practical ethics by showing how desires, anger, ambition, and vanity distort action unless rectified through knowledge, discipline, and remembrance. Taha Hashas, in modern Islamic philosophy, likewise argues that responsibility cannot be reduced to procedural obligation, because the human being is a trusteeship-bearing moral subject whose action is inseparable from inward answerability. Across these thinkers, the pattern is consistent: moral order in society depends ultimately upon the moral ordering of the self (Hashas, 2015).

This section therefore yields a clear conclusion. Hadīth, *aqwal*, and *athar* do not treat morality as a system sustained primarily by external enforcement. They construct a pedagogy of inner life in which intention, heart, silence, fasting, vigilance, and self-accounting become the means by which *taqwa* is stabilized in practice. The result is a model of moral formation that is highly relevant to the age of AI. Where external governance can supervise, the Prophetic and early Islamic tradition seeks to transform. Where institutions can audit, this tradition trains the believer to audit himself. Where technical systems may increase the reach of control, Hadīth and *athar* remind us that the decisive ethical question remains whether the human being has become capable of governing himself before Allah.

TOWARD AN ISLAMIC FRAMEWORK FOR AI ETHICS

If external governance is necessary yet insufficient, and if *taqwa* names the inward authority without which responsible action remains unstable, then the next task is constructive: to identify a framework through which artificial intelligence may be ethically integrated without being morally enthroned. The governing principle of such a framework is simple but decisive. Technology may assist, extend, and support human action, but it must not occupy the place of moral leadership. In Islamic terms, the human being is not permitted to surrender the burden of moral direction to the instrument he has made, because he remains the bearer of trust, responsibility, and answerability before Allah. The ethical question, therefore, is not whether AI may be used, but under what ordering of authority it may be used rightly (*ibid*).

The first foundation of this framework is *amanah*. The Qur'an declares:

إنا عرضنا الأمانة على السماوات والأرض والجبال فأبين أن يحملنها
وأشفقن منها وحملها الإنسان إنه كان ظلوما جهولا

“Indeed, We offered the Trust to the heavens and the earth and the mountains, but they declined to bear it and feared it; yet man undertook it. Indeed, he has always been prone to injustice and ignorance” (Haleem, 2004).

This verse establishes a basic moral orientation for all human action, including technological action. Human beings are not absolute owners of power but entrusted bearers of responsibility. When applied to AI, *amanah* requires that systems be designed, deployed, and governed as instruments under trust rather than as autonomous replacements for human conscience. This changes the ethical grammar of technology. It means that capability alone does not justify use, efficiency alone does not determine legitimacy, and innovation alone does not define progress. The relevant question becomes whether a given technological intervention honors the trust borne by the human being or weakens his sense of answerability (Azram, 2012).

The second foundation is *maqasid al-shariah*. A constructive Islamic AI ethics cannot remain at the level of general piety; it must possess evaluative criteria. The classical *maqasid* tradition, especially as systematized by al-Shatibi, offers precisely such a structure through the preservation of religion, life, intellect, lineage, and property. In contemporary moral reasoning, these objectives also illuminate adjacent concerns such as dignity, justice, and public welfare. Their significance for AI ethics is considerable. A system that degrades intellect through habitual overdependence, manipulates desire for commercial gain, obscures accountability in matters affecting life and livelihood, or undermines human dignity through invasive surveillance cannot be treated as ethically neutral merely because it is technically functional. *Maqasid* supplies a disciplined standard

through which benefit and harm may be evaluated beyond raw utility or market success (Al-Shatibi, 2003).

The preservation of intellect is particularly important in the present context. Many contemporary AI systems are celebrated for reducing cognitive burden, increasing speed, and simplifying decision pathways. These benefits are real, but they introduce an ethical tension. A system that assists judgement may be beneficial; a system that gradually replaces the habit of judgement threatens the maqasid-related protection of aql. The issue here is not merely whether the output is correct, but whether the user remains intellectually and morally capable of evaluating, resisting, refining, and taking responsibility for that output. A maqasid-based framework therefore requires that AI augment human judgement without eroding the very faculties that make judgement possible (Kim, 2025). A third foundation is niyyah. Islamic ethics has always insisted that the moral worth of action cannot be separated from intention. This principle must be brought into technological ethics more explicitly than most contemporary governance models allow. Systems are often evaluated through performance, safety metrics, documented compliance, or legal conformity, all of which matter. Yet Islamic moral reasoning asks an additional question: for what purpose is the system being built, and to what end is it being deployed? A design process governed by niyyah would require that developers, institutions, and regulators articulate the moral purpose of AI systems in advance, rather than treating purpose as an afterthought once capacity has already been achieved. This does not replace technical testing; it precedes it by ordering technical work toward morally intelligible ends. A fourth foundation is the principle of non-harm. The Prophetic maxim states:

لا ضرر ولا ضرار

“There should be neither harm nor reciprocal harm”. (Anas, 2005).

This maxim is exceptionally important for AI ethics because it moves the discussion from abstract legitimacy to concrete consequences. Systems that intensify addiction, facilitate manipulation, normalize unjust surveillance, amplify bias, or distribute serious harms without clear accountability fall directly within the field of *darar*. An Islamic framework cannot be satisfied with usefulness in one domain if significant moral injury is displaced into another. The prohibition of harm therefore requires more than post hoc repair; it demands anticipatory ethical scrutiny of foreseeable injuries to dignity, intellect, justice, and social trust (Anas, 2005). A fifth foundation is the distinction between human oversight and moral outsourcing. Much contemporary AI policy language rightly emphasizes keeping a human in the loop, on the loop, or over the loop. Yet from an Islamic ethical perspective, this remains inadequate unless the human agent retains substantive moral authorship. A nominal human presence does not solve the ethical

problem if the human has become dependent, uncritical, or unwilling to overrule technically efficient but morally unsound outcomes. The real question is not whether a human signs off at the end, but whether that human remains morally awake, intellectually capable, and conscience-led throughout the process. In this sense, oversight must mean more than procedural involvement; it must mean preserved human responsibility (Dignum, 2019).

Muslim thinkers strengthen this point by refusing to separate outer systems from the inner state of the acting person. Azram’s discussion of governance is especially useful because it insists that scientific and professional capacity must remain under the guidance of revelation and that power, when divorced from wisdom and God-consciousness, becomes corruptive and ruinous. This argument directly supports the central principle of the present article: technological strength is not rejected, but it must remain subordinate to morally formed agency. In other words, *quwwah* is valuable only when governed by *taqwa*. Once that order is inverted, the instrument begins to dominate the one who was meant to guide it (Azram, 2012).

This same priority appears in Muslim reflections on the *taqwa-quwwah* relationship. The point is not that material strength, science, organization, or strategic capacity are irrelevant. Rather, the warning is that development models that privilege *quwwah* while marginalizing *Taqwa* produce imbalance in attitudes, motivation, and morality. Such imbalance is especially dangerous in the technological domain, because it can make power appear self-justifying. An Islamic AI ethics must therefore reject both anti-technical romanticism and morally ungoverned technicism. It affirms advancement, but only under the prior rule of conscience (Hamid, 2003).

From these foundations, several practical guardrails follow. AI should be used to assist human deliberation, not to erase it. It should reduce burdens without producing moral passivity. It should expand access to knowledge without corroding intellectual effort. It should serve justice and welfare without exploiting vulnerability, desire, or heedlessness. It should be deployed in high-impact domains only where accountability remains reviewable, contestable, and humanly owned. Most importantly, no technological process should be treated as morally self-sufficient. Every system remains subordinate to the judgement of those who build, authorize, and use it, and those persons remain answerable before Allah for the ends they choose and the harms they permit (UNESCO, 2021).

The constructive force of this framework lies precisely in its balance. It does not call for the abandonment of AI, nor for a retreat from science, technical skill, or institutional governance. It affirms beneficial innovation, but refuses to let innovation

become a new sovereign. It welcomes technological assistance, but denies technology the right to moral command. It accepts regulation, but insists that regulation alone cannot generate conscience. In that sense, the framework may be summarized in one governing proposition: AI must remain under conscience, because conscience alone can answer for the use of AI. The next section therefore addresses the most likely objections to this claim and clarifies its scope more fully.

OBJECTIONS AND RESPONSES

A serious article in AI ethics cannot proceed by assertion alone. If the argument of this study is that external governance remains necessary yet insufficient, and that AI must remain under the leadership of conscience rather than occupy the driver's seat itself, then several objections arise naturally. Some come from contemporary governance discourse, some from secular ethical theory, and some from within the very logic of moral anthropology. Addressing them is essential, because the aim of this article is not to oppose technological development, but to clarify the conditions under which development remains genuinely beneficial and morally ordered (Dignum, 2019).

A first objection states that regulation and governance were never meant to create virtue in the first place. Their function, it may be argued, is limited and practical: to reduce harm, distribute responsibility, and establish institutional safeguards. From this perspective, it is unfair to criticize AI governance for not producing inward moral excellence, because that was never its task. This objection carries real force and must be granted in part. External governance is indeed designed primarily to restrain, coordinate, and supervise. The present article does not deny that purpose. Its argument is narrower. It holds that when societies begin to behave as though compliance mechanisms are morally exhaustive, they risk neglecting the interior conditions without which compliance itself becomes shallow, selective, or strategically performative. The claim, therefore, is not that governance has failed at being governance, but that governance cannot bear the whole burden of ethics by itself (UNESCO, 2021).

A second objection argues that secular moral philosophy already possesses robust accounts of character, virtue, and moral formation, so the present article overstates the gap in contemporary ethical discourse. This objection is also important. It is true that virtue ethics, moral psychology, and traditions of character formation have long addressed the relation between outward action and inward disposition. MacIntyre's recovery of virtue, for example, was itself a protest against the fragmentation of moral life into procedural remnants. The response, however, is that contemporary AI governance discourse has been driven

more strongly by principles of safety, fairness, accountability, transparency, and rights than by thick accounts of moral formation. The issue is not the total absence of virtue ethics in modern thought, but its relative marginality within dominant institutional frameworks for AI. The article therefore does not deny secular resources; it argues that current governance practice remains thinner than the moral problem it seeks to solve (MacIntyre, 2007).

A third objection states that internal conscience is itself unreliable. Human beings rationalize, deceive themselves, misname desire as sincerity, and often invoke inward conviction in order to escape accountability. If this is so, then grounding AI ethics in conscience or *taqwa* may appear dangerously subjective. This objection is perhaps the most serious, and Islamic moral thought has never ignored it. On the contrary, the classical tradition begins precisely from the instability of the *nafs*. That is why *taqwa* is not defined as instinctive inward feeling, but as disciplined moral formation under revelation, worship, remembrance, self-accounting, and the correction of the heart. Conscience in the sense defended here is not autonomous self-certainty; it is a trained inward authority shaped by divine guidance and sustained through *muhasabah* and *muraqabah*. The answer, therefore, is not to abandon inward authority because it can be corrupted, but to distinguish formed conscience from ungoverned impulse (Al-Muhasibi, 1964).

A fourth objection maintains that religious categories such as *taqwa*, *amanah*, and *ihsan* are not universal enough for a global discussion of AI ethics. Since AI governance operates across plural societies, legal systems, and moral traditions, an explicitly Islamic framework may be thought too particular to offer general relevance. This objection should be met with clarity. The article does not claim that all participants in global AI ethics must share Islamic metaphysical commitments in order to understand its argument. Rather, it offers an Islamic normative intervention into a global question. Such an intervention is legitimate for two reasons: first, because all serious ethical traditions contribute from within their own conceptual vocabularies; and second, because the specific problem addressed here, namely the insufficiency of external governance without inward moral formation, is intelligible even to those who do not share the theological premises of Islam. The universality claimed here is therefore analytical before it is confessional (Izutsu, 2002).

A fifth objection argues that the benefits of AI are too great, and the competitive pace of innovation too intense, for societies to burden technological development with thick moral requirements. In this view, excessive ethical hesitation risks slowing beneficial innovation in medicine, education, accessibility, governance, and economic productivity. The response of this article is not to deny benefit, but to reject the assumption that speed

and benefit are sufficient grounds of legitimacy. History repeatedly shows that human beings can achieve extraordinary technical capacity while remaining morally unprepared for its consequences. The issue, then, is not whether AI can do good, but whether societies possess the moral discipline to govern its power without surrendering to convenience, domination, or dependency. A technology that assists without displacing judgement is a benefit; a technology that gradually erodes the authority of conscience while increasing operational ease may deliver immediate gains at deeper long-term cost (Ellul, 1964).

A sixth objection concerns human oversight. Many policy frameworks already insist that humans remain in the loop, on the loop, or over the loop. From this perspective, the article may appear to demand something already secured in contemporary governance language. Yet this objection rests on an ambiguity. Human involvement is not identical with morally serious oversight. A human being may formally approve an AI-mediated process while remaining intellectually passive, institutionally pressured, or morally disengaged. In that case, the “human in the loop” becomes little more than a symbolic remainder rather than a substantive moral agent. The point of the present article is that oversight becomes ethically meaningful only when the human participant retains real judgement, real responsibility, and real freedom to refuse technically efficient but morally unsound outcomes. A merely ceremonial human presence does not resolve the ethical problem (NIST, 2024).

A seventh objection may come from the opposite direction and claim that the article underestimates the extent to which external systems are indispensable precisely because human beings are morally weak. If corruption, impulsiveness, and injustice are permanent features of human life, then strong procedural control may seem more realistic than any hope for widespread moral formation. This objection contains an important realism, and Islamic thought would not dismiss it. The Shariah itself includes outward regulation, evidentiary procedure, institutional order, and visible accountability because human beings cannot be trusted without structure. Yet Islamic ethics never concludes from human weakness that inward reform is irrelevant. Rather, it insists on the joint necessity of outer rule and inner discipline. The present article therefore does not replace governance with spirituality; it argues that governance without inward authority is structurally incomplete, while inward aspiration without governance remains institutionally fragile (Al-Shatibi, 2003).

An eighth objection suggests that the thesis of “AI under conscience” risks romanticizing the human being and demonizing the tool. This concern should also be answered directly. The article does not imagine the human agent as morally pure, nor the machine as metaphysically corrupt. Its claim is functional and ethical, not mythic. Human beings remain responsible because

they alone bear intention, accountability, and answerability in the full moral sense. Tools, however sophisticated, do not become rightful bearers of moral leadership merely because they become more capable. The distinction between leader and assistant is therefore not based on technological fear, but on moral ontology. AI may extend capacity, but it cannot inherit the burden of answerability that belongs to the human being as bearer of Amanah (Hashas, 2015).

CONCLUSION

This article has argued that the core limitation of contemporary AI ethics is not the absence of regulation, but the assumption that regulation alone can secure moral order. External governance can impose safeguards, reduce risk, improve accountability, and regulate conduct. It cannot by itself produce sincerity, restraint, justice, self-accounting, or inward responsibility. For that reason, compliance is necessary, but never sufficient. The central claim of this study is precise: AI must remain an assistant to conscience-led human agency, not its driver. Once technology begins to lead and the human being becomes morally passive, dependent, or merely supervisory in name, the ethical order is inverted. In such a condition, technical power may increase while moral responsibility weakens. That inversion is the real danger. The Islamic tradition offers a more complete foundation for safety because it does not confine ethics to external control. Through taqwa, amanah, ihsan, niyyah, muraqabah, muhasabah, and maqasid al-shariah, Islam secures both the outer and inner dimensions of human action. It regulates conduct, but also disciplines intention. It establishes limits, but also forms conscience. It addresses public justice, private restraint, intellectual integrity, and accountability before Allah. This makes Islamic principles uniquely comprehensive in ensuring safety across all respects: personal, social, intellectual, institutional, and civilizational. The conclusion is therefore categorical. AI may be used, developed, and advanced for genuine human benefit. But moral authority must remain with the human being under divine guidance. The safest future will not be produced by code alone. It will be secured when technology serves, conscience leads, and ethical order remains grounded in the enduring authenticity and completeness of Islamic moral principles.

RECOMMENDATIONS

1. Legally affirm human moral primacy. Every AI policy should state that final moral and legal responsibility remains with accountable human persons and institutions, never with the system.
2. Restrict AI to an assistive role.

AI should be authorized to support analysis, prediction, drafting, and coordination, but not to become the final driver in morally consequential decisions.

3. Mandate meaningful human oversight.

“Human in the loop” must mean real authority to review, question, suspend, and override AI outputs, not mere formal presence.

4. Classify high-risk moral domains.

In areas such as justice, policing, warfare, healthcare, education, child-related matters, religion, and public governance, AI use must face stricter approval and supervision.

5. Require intention-and-purpose review before deployment.

No system should be approved only on technical performance; regulators must ask why it is being used, for whose benefit, and with what foreseeable moral effects.

6. Protect human intellect and judgement.

Policy must prevent overdependence by requiring that AI systems augment rather than erode reasoning, deliberation, memory, and professional competence.

7. Make harm prevention a continuing duty.

Systems must be monitored for bias, manipulation, addiction, privacy invasion, dehumanization, deskilling, and unjust exclusion throughout their lifecycle, not only before launch.

8. Build traceability and answerability into governance. Every significant AI-assisted decision must remain attributable, reviewable, contestable, and auditable by competent human authorities.

9. Require ethical training for developers and users.

Technical competence alone is insufficient; institutions should include structured training in responsibility, restraint, accountability, and public trust.

10. Adopt a non-substitution rule.

Any AI application that effectively transfers moral leadership from conscience-led humans to machine-led processes should be restricted, redesigned, or rejected.

AUTHOR CONTRIBUTIONS

MFR contributed to conceptualization, methodology design, data curation, and drafting of the manuscript. MK performed analysis, and validated the results.

CONFLICTS OF INTEREST

No conflict of interest among the authors to declare

DATA AVAILABILITY

Data will be made available on a fair request to the corresponding author

ETHICS APPROVAL

Not applicable to this paper.

FUNDING SOURCE

Self-funded.

REFERENCES

- Abdel Haleem MAS (2004) *The Qur'an*. Oxford: Oxford University Press.
- Al-Asfahani al-Raghib (n.d.) *Al-Mufradat fi Gharib al-Qur'an*. Beirut: Dar al-Ma'rifah.
- Al-Bukhari MI (n.d.) *Sahih al-Bukhari*. Cairo: Dar Ibn Kathir/Dar Tawq al-Najah.
- Al-Ghazali AH (1993) *Ihya' 'Ulum al-Din*. Beirut: Dar al-Ma'rifah.
- Al-Hadith al-Nabawi.
- Al-Makki AT (1997) *Qut al-Qulub*. Beirut: Dar Sadir.
- Al-Muhasibi al-Harith ibn Asad (1964) *Risalat al-Mustarshidin*. Beirut: Dar al-Kutub al-'Ilmiyyah.
- Al-Shatibi AI (n.d.) *Al-Muwafaqat fi Usul al-Shari'ah*. Beirut: Dar al-Ma'rifah.
- Al-Tabari AJMIJ (2003) *Jami' al-Bayan 'an Ta'wil Ay al-Qur'an*. Riyadh: Dar 'Alam al-Kutub.
- Azram M (2012) Principles of Good Governance in Islam. In: *Management of Resources in Muslim Countries and Communities: Challenges and Prospects*. Kuala Lumpur: IIUM Press.
- Dignum V (2019) *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham: Springer. <https://doi.org/10.1007/978-3-030-30371-6>.
- Ellul J (1964) *The Technological Society*. New York, NY: Vintage Books.
- Floridi L and Cowls J (2019) A unified framework of five principles for AI in society. *Harvard Data Science Review* 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>.
- Foucault M (1995) *Discipline and Punish: The Birth of the Prison*. Translated by A Sheridan. New York, NY: Vintage Books.
- Hamid AFA (2003) The Taqwa versus Quwwah dichotomy: An Islamic critique of development via the Malaysian Bumiputera policy. *Kajian Malaysia: Journal of Malaysian Studies* 21(1-2): 123–162.
- Heidegger M (1977) *The Question Concerning Technology and Other Essays*. Translated by W Lovitt. New York, NY: Harper & Row.
- Hashas M (2015) Taha Abderrahmane's trusteeship paradigm: Spiritual modernity and the Islamic contribution to the formation of a renewed universal civilization of ethos. *Oriente Moderno* 95(1-2): 67–105. <https://doi.org/10.1163/22138617-12340077>.
- Kim J (2025) Perceptions and preparedness of K-12 educators in adopting generative AI. *Research in Learning Technology* 33: 3448. <https://doi.org/10.25304/rlt.v33.3448>.
- Ibn Kathir IIU (1999) *Tafsir al-Qur'an al-'Azim*. Riyadh: Dar Taybah.
- Ibn Manzur (1997) *Lisan al-'Arab*. Beirut: Dar Sadir.
- Ibn Rajab al-Hanbali (n.d.) *Lata'if al-Ma'arif*. Beirut: Dar Ibn Hazm.
- Izutsu T (2002) *Ethico-Religious Concepts in the Qur'an*. Montreal and Kingston: McGill-Queen's University Press.
- MacIntyre A (2007) *After Virtue: A Study in Moral Theory*. 3rd edn. Notre Dame, IN: University of Notre Dame Press.
- Malik ibn Anas (2005) *Al-Muwatta'*. Beirut: Dar al-Kutub al-'Ilmiyyah.
- NIST (2024) *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. Gaithersburg, MD: National Institute of Standards and Technology. Available at: <https://www.nist.gov/>.
- OECD (2026) *The Agentic AI Landscape and Its Conceptual Foundations*. Paris: Organisation for Economic Co-operation and Development. https://www.oecd.org/en/publications/the-agentic-ai-landscape-and-its-conceptual-foundations_3f78135c-en.html.
- UNESCO (2021) *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO. Available at: <https://www.unesco.org/>.
- Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York, NY: PublicAffairs.