Toward an Islamic Ethics and Fiqh of Artificial Intelligence

MOHAMED ABUTALEB
MOHAMMED ANSARI
KENAN ALKIEK
SULEIMAN HANI
UMER KHAN

yaqeen™

Author Biography

Dr. Mohamed AbuTaleb serves as VP of Research and Content Strategy at Yaqeen Institute for Islamic Research, faculty and former dean at several graduate seminaries, and Resident Scholar in the Research Triangle Region of North Carolina, USA. His research interests include Islamic institutional excellence, ethics of artificial intelligence (AI) and technology, and community leadership and resilience.

Dr. Mohammed Ansari holds a PhD in computer sciences from the University of Wisconsin-Madison with a specialization in artificial intelligence. He holds a senior scientist role in the industry and is passionate about developing technological solutions that rise to address contemporary challenges for the *umma*.

Kenan Alkiek holds degrees in computer science and applied statistics and is currently a PhD candidate at the University of Michigan. His research focuses on reasoning and reliability in language models, with broader interests in artificial intelligence.

Dr. Suleiman Hani serves as Dean of Academic Affairs at AlMaghrib Institute and Research Fellow at Yaqeen Institute. A distinguished scholar and international lecturer, he specializes in Islamic studies, philosophy, leadership, and human rights.

Shaykh Umer Khan serves as an executive member of the Fiqh Council of North America, teaches the Islamic sciences at the Institute of Knowledge and Al-Salam Institute, researches and writes fatwas (legal verdicts) for various institutions, and serves as a sharia advisor to a number of companies. He holds degrees and professional qualifications in technology, Islamic finance, and Islamic law.

Disclaimer: *The views, opinions, findings, and conclusions expressed in these papers and articles are strictly those of the authors. Furthermore, Yaqeen does not endorse any of the personal views of the authors on any platform. Our team is diverse on all fronts, allowing for constant, enriching dialogue that helps us produce high-quality research.*

Introduction

Artificial intelligence (AI) has moved rapidly from the realm of futuristic speculation to an indispensable part of modern life. Embedded in everyday technologies, AI now fundamentally shapes how we communicate, work, and interact with the world. The scale, speed, and reach of AI's integration into society mark it as a uniquely disruptive force that brings immense promise but also unprecedented ethical, social, and cultural challenges.

Navigating this disruption requires moving beyond reductive binaries of uncritical technological optimism or fatalistic pessimism. The ethical challenges of AI are rarely simple; they frequently present a deeply complex matrix where benefits and harms are inextricably linked. To address this, we propose that the Islamic ethical lens is essential for evaluating AI use and development. Capable of providing a proactive, driving framework that informs innovation and technology, it offers values that help balance individual and communal considerations. This framework goes beyond downstream mitigation to proactively judge and shape form, substance, and approach.[1]

The paper begins by surveying the emergence and projected impact of AI, alongside its inevitable adoption across a myriad of contexts. By exploring specific applications and integrating relevant guidelines from Islamic scholars, we articulate how core ethical principles can be operationalized. After situating the importance of clear rulings and principles assigned to the technology domain, we present the jurisprudence of balancing (*fiqh al-muwazaat*), a sophisticated methodology that provides a structured approach to ethical dilemmas where utility and risk coexist.

This paper is methodologically oriented, offering general principles that one should take into account in their ethical decision-making. Generating definitive rulings for every AI case is beyond the scope of this paper; many of the issues raised here

[1] It is important to note that Islamic ethics is not identical to *fiqh* nor an extension of it; rather, *fiqh* is a partial legal articulation of the wider moral vision found in Islamic ethics. While scholars treat *fiqh* as encompassing moral reasoning, reducing ethics to law misses important dimensions like intention, character, and spiritual excellence.

remain open questions requiring interdisciplinary *ijtihad*. Ultimately, the framework presented here establishes the theoretical and practical necessity of an Islamic approach. It provides an approach for the ongoing work required to assess these technologies and the profound civilizational stakes of the algorithmic age.

Why technology needs ethics

Social media offers a clear precedent for the risks of uncritical technological adoption: Society and Islamic scholarship were slow to assess its ethical implications, allowing harms to become deeply embedded before meaningful safeguards emerged. What began as niche platforms rapidly expanded into a global infrastructure shaping communication, culture, and economics, driven largely by companies prioritizing growth, engagement, and profit over long-term societal well-being, while regulatory and ethical responses lagged behind. The Muslim *ummah*'s response was similarly delayed and reactive, with guidance emerging only after widespread adoption had already formed habits and constrained meaningful intervention; even then, Islamic ethical engagement largely focused on mitigating harms within an assumed inevitability rather than issuing foundational judgments about the technology's purpose, means, and anticipated consequences. Today, generative AI presents a parallel but higher-stakes transformation, advancing more quickly and concentrating unprecedented power in a small number of profit-driven technology firms. Without proactive, principled evaluation of its underlying purposes, design, and consequences, there is a serious risk of repeating the same pattern, allowing structural harms to take root before they are fully understood, underscoring the urgent need for ethical frameworks to guide technological development from the outset rather than attempting to mitigate damage after the fact.

AI is transforming communication, education, health care, finance, security, and culture at once. However, we know that it will not always be beneficial. The same systems that can expand access to knowledge, improve medical diagnosis, streamline services, and unlock new forms of creativity can also entrench bias, destabilize trust, displace livelihoods, and concentrate power. Human design choices, from the way an algorithm prioritizes information to features that

encourage user engagement, reflect the values and assumptions of its creators. These choices may appear technical, but they carry ethical weight because they shape how people interact, what information they encounter, and which behaviors are rewarded or discouraged.

Compounding this challenge, technology often spreads at a pace much faster than laws, cultural norms, or institutions can adjust. Regulation typically follows innovation rather than guiding it, which means that once a technology is entrenched, its harms are much harder to undo. In moments like this, societies often let profit and speed set the terms while communities absorb the costs. For this reason, ethical foresight is not optional but essential: Societies must grapple with potential consequences in advance, rather than waiting until negative impacts have already become embedded.

One of the clearest examples of how technology is governed in ways that prioritize growth over responsibility is Section 230 of the Communications Decency Act, passed in 1996. Often described as "the twenty-six words that created the internet," [2] this single provision not only enabled the rapid expansion of online platforms but also institutionalized a model of governance in which legal immunity outpaces ethical accountability. By granting platforms immunity from liability for user-generated content, while allowing them to moderate in "good faith," the law treats companies such as Facebook, X (formerly Twitter), and YouTube as distributors rather than publishers, shielding them from responsibility for most of the content they host.

By removing the threat of liability, Section 230 helped create the conditions under which platforms could maximize growth and user engagement while minimizing attention to potential harms. Once user-generated content became the lifeblood of the internet, algorithms were optimized to capture attention and drive profits, not to safeguard communities. This model has been described as "surveillance capitalism," a new economic order in which human experience is treated as free

[2] Jeff Kosseff, *The Twenty-Six Words That Created the Internet* (Cornell University Press, 2019).

raw material for hidden commercial practices of extraction, prediction, and sales.[3] Its harms are not only societal but deeply personal, often disproportionately borne by vulnerable populations.[4] For example, content moderation and AI model training are frequently outsourced to workers in developing countries, who must process trauma-inducing content under precarious conditions, often without adequate psychological support. In this way, technology companies generate immense private wealth while externalizing social and human costs—a pattern consistent with broader dynamics in modern capitalism, but amplified by the scale and reach of digital platforms.[5]

This legal framework can be clarified through an analogy: A publisher, such as a newspaper, is liable for the content it prints, while a distributor, such as a bookstore, is not responsible for the material it sells. By classifying platforms as distributors, Section 230 created a system in which companies can profit from harmful content without bearing corresponding responsibility. This reveals a fundamental limitation of secular legal frameworks: They determine liability but often remain silent on moral obligation. Legal immunity, in this sense, does not equate to ethical innocence. It is precisely this gap, between what is legally permissible and what is ethically justifiable, that invites alternative approaches to governance. Islamic ethics, with its emphasis on accountability, justice, and the moral consequences of action, offers a more comprehensive framework for addressing the responsibilities of technological power.

An Islamic ethical governance approach

For Muslims, Islamic ethics serves as a proactive, spiritually anchored framework that governs not only the application of a technology, but its very conception. This multitiered paradigm is not a novel construction; rather, it is a direct reflection of the divine mandate established in the Holy Qur'an and the Prophetic Sunnah. The Qur'an explicitly commands a comprehensive moral framework that transcends

[3] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).
[4] Madhumita Murgia, *Code Dependent: Living in the Shadow of AI* (Henry Holt and Company, 2024).
[5] Christopher Marquis, *The Profiteers: How Business Privatizes Profits and Socializes Costs* (PublicAffairs, 2024).

bare minimum compliance, stating: "Indeed, Allah orders justice [*ʿadl*], good conduct [*ihsan*], and generosity to close relatives."[6] In the realm of technological innovation, *ʿadl* necessitates strict adherence to legal boundaries that prevent harm and uphold human rights, while *ihsan* demands that the innovator's internal disposition and the technology's ultimate trajectory actively seek the highest communal good. Furthermore, the Prophet ﷺ established that human agency and power constitute a profound trust in his statement, "Every one of you is a shepherd and is responsible for his flock."[7] When developers and executives design AI tools, they wield systems capable of shaping the cognitive and moral fabric of billions, thus assuming the role of shepherds over a vast digital expanse. Consequently, their standard of accountability cannot be limited to merely dodging secular legal liability; it must encompass the holistic fulfillment of this divine trust.

Unlike many contemporary secular models, this Islamic paradigm operates at the very genesis of technological development.[8] By judging a technology's form, purpose, and design from its inception, Islamic jurisprudence provides a mechanism for critiquing systems that are fundamentally corrupt by design, recognizing that structural harms cannot be adequately resolved through superficial, downstream mitigation.[9] Furthermore, this paradigm radically expands the calculus of benefit (*maslaha*) and harm (*mafsada*). It extends the scope of assessment beyond immediate worldly utility to encompass consequences in both this life and the hereafter. This dual-horizon accountability provides a robust defense against vectors of unregulated profit and exploitation that often evade human legal enforcement, anchoring the innovator's restraint in the ultimate reality of divine oversight.[10]

---

[6] Qur'an 16:90.

[7] *Ṣaḥīḥ al-Bukhārī*, *Kitāb al-aḥkām*, no. 7138.

[8] This proactive approach aligns with the foundational Islamic emphasis on intention (*niyya*) preceding action. The Prophet ﷺ stated, "Indeed, actions are judged by intentions." *Ṣaḥīḥ al-Bukhārī*, *Kitāb badʾ al-waḥy*, no. 1. In the context of technology, the initial purpose and design carry as much ethical weight as the final output.

[9] This reflects the jurisprudential principle of *sadd al-dharāʾiʿ* (blocking the means to harm), which mandates addressing the root causes of corruption before they materialize and is discussed further later in this paper. See Abū Isḥāq al-Shāṭibī, *al-Muwāfaqāt fī uṣūl al-sharīʿa* (Dār al-Maʿrifa, 1997), 4:198–200. It also aligns with the legal maxim that what is built upon an invalid foundation is itself invalid (*mā buniya ʿalā bāṭil fa-huwa bāṭil*).

[10] The Sharia uniquely defines human welfare (*maṣlaḥa*) as encompassing both worldly prosperity and salvation in the next life. See ʿIzz al-Dīn Ibn ʿAbd al-Salām, *Qawāʿid al-aḥkām fī maṣāliḥ al-anām* (Maktabat al-Kulliyyāt al-Azhariyya, 1991), 1:11. The Qur'an continually reinforces this ultimate accountability that transcends human

At its foundation, justice (*ʿadl*) is operationalized through *ahkam*—clear legal rulings that establish nonnegotiable boundaries of permissibility (halal) and prohibition (haram). In a technological landscape driven by speed and competition, these serve as definitive guardrails, ensuring that core principles such as the preservation of intellect, dignity, and privacy are not sacrificed for innovation. These limits, or red lines, understood as the divinely ordained boundaries (*hudud Allah*), govern all human activity, including the development of AI.[11]

Yet the Sharia is not merely a restrictive code; it is a visionary system designed to promote human welfare and flourishing. To construct a functional Islamic ethics of artificial intelligence, we must trace a clear, teleological trajectory from universal divine intent down to practical engineering. This framework begins with the *maqasid ʿamma*—the overarching, constitutional objectives of the Sharia, such as the preservation of human intellect (*hifz al-ʿaql*) and human dignity.[12] However, because these universal aims are too broad to directly regulate complex algorithms, scholars and technologists must collaboratively derive *maqasid khassa*: objectives specific to a single domain of activity—here, the digital realm.[13] In the context of generative technology, a *maqsad khass* might, for example, mandate that AI systems function as transparent intellectual scaffolding rather than deceptive, autonomous replacements for human moral reasoning. These domain objectives are in turn specified as *maqasid juzʾiyya*: the concrete intents behind particular design rules, such as requiring a system to disclose its limitations or to route personal religious questions to a qualified human scholar. This entire hierarchy results in

courts: for example, "So whoever does an atom's weight of good will see it, and whoever does an atom's weight of evil will see it" (Qur'an 99:7–8).

[11] The concept of immutable red lines is rooted in the Qur'anic concept of *ḥudūd Allāh* (the limits of Allah). The Qur'an warns, "These are the limits of Allah, so do not transgress them. And whoever transgresses the limits of Allah—it is those who are the wrongdoers" (Qur'an 2:229). Applying this to technology ensures that innovation remains constrained by divine boundaries.

[12] Abū Ḥāmid al-Ghazālī is traditionally credited with crystallizing the five universal necessities (*al-ḍarūriyyāt al-khams*), which include the preservation of intellect. See Abū Ḥāmid al-Ghazālī, *al-Mustaṣfā fī ʿilm al-uṣūl* (Dār al-Kutub al-ʿIlmiyya, 1993), 1:174. For a contemporary expansion on human dignity as a universal objective, see Mohammad Hashim Kamali, *Maqāṣid al-Sharīʿah Made Simple* (International Institute of Islamic Thought, 2008), 22–24.

[13] The conceptualization of domain-specific objectives was heavily developed by the influential Maliki jurist Ibn ʿĀshūr, who argued that broad legal philosophy must be specialized to effectively govern distinct fields of human endeavor. See Muhammad al-Tahir ibn Ashur, *Treatise on Maqāṣid al-Shariʿah*, trans. Mohamed El-Tahir El-Mesawi (International Institute of Islamic Thought, 2006), 183–87.

maslaha: Tangible public benefit is secured and harm is repelled when accessible, bias-mitigated tools are deployed to solve pressing societal challenges.[14] This progression from universal principle, to domain objective, to the intent of a specific rule, and finally the benefit it secures, ensures that innovation remains anchored in the pursuit of human flourishing.

Yet external structures alone are insufficient. Islamic governance also demands the cultivation of *adab*—the spiritual etiquette, ethical character, and professional virtues required of those who design and deploy technology.[15] Without this moral foundation, even the most robust legal and institutional safeguards will fail. Engineers and executives must recognize their work as a form of stewardship (*khilafa*), in which technical decisions carry spiritual and societal consequences. Anchored by *tarbiya* (moral discipline), this internal ethic bridges the gap between abstract code and real-world impact, ensuring that responsibility is imposed from within.

To enforce meaningful ethical guardrails in AI, mechanisms of liability must also accompany moral principles. The Islamic tradition offers a clear precedent in its treatment of medical ethics (*adab al-tabib*), where practitioners were entrusted with human well-being and held liable for foreseeable harm.[16] Developers of generative AI occupy a comparable role today, yet the current landscape—shaped by competitive pressures and exemplified by private AI start-up Anthropic's retreat from unilateral safety constraints[17]—disincentivizes restraint. Islamic jurisprudence

---

[14] For a foundational discussion on how the Sharia secures tangible public welfare (*maṣlaḥa*) and repels harm in practical application, see al-Shāṭibī, *al-Muwāfaqāt fī uṣūl al-sharīʿa*, 2:9–12. For a modern systems approach to applying these concepts to contemporary policy and technology, see Jasser Auda, *Maqasid Al-Shariah as Philosophy of Islamic Law: A Systems Approach* (International Institute of Islamic Thought, 2008), 22–25.

[15] The Prophet ﷺ emphasized that internal character is the heaviest weight on the scales of judgment. See *Sunan al-Tirmidhī*, *Kitāb al-birr wa-l-ṣila*, no. 2002.

[16] The Prophet ﷺ stated: "Any physician who practices medicine when he was not previously known to be a practitioner and causes harm [to the patient] will be held liable." *Sunan Abī Dāwūd,* no. 4587; *Sunan al-Nasāʾī*, no. 4830; *Sunan Ibn Māja*, no. 3466. The hadith has been transmitted through various chains and many scholars accept it as a valid basis for legal rulings. Among contemporary scholars, al-Albānī graded it as *ḥasan* and al-Arnāʾūṭ graded it as *ḥasan li-ghayrih*. Some scholars consider the narration as *mursal*. The broader concept of a practitioner's responsibility is well established in related narrations outside of the specific medical contextual application.

[17] Billy Perrigo, "Exclusive: Anthropic Drops Flagship Safety Pledge," *Time*, February 24, 2026, https://time.com/7380854/exclusive-anthropic-drops-flagship-safety-pledge/; see also "Responsible Scaling Policy Version 3.0," Anthropic, February 24, 2026, https://www.anthropic.com/news/responsible-scaling-policy-v3.

addresses this through a dual standard of accountability: responsibility before God and enforceable liability in this life (*qada'*). Rooted in Prophetic guidance, this framework obliges practitioners to anticipate risks, implement safeguards, and provide restitution when harm occurs. Applied to AI, it challenges liability-shielded models by demanding accountability for foreseeable harms and embedding ethical responsibility into both intention and action, ensuring that innovation is tempered by the protection of human well-being.

Further details on the Islamic standards of liability in AI applications are explored in a later section and by reference. To help contextualize the coming frameworks, we also provide an accessible technical primer geared on demystifying generative AI in the appendix. We now move to outline the framework of *fiqh al-muwazanat* (jurisprudence of balancing) and its proposed application to AI ethics. This framework is anchored in concrete legal rulings (*ahkam*) such that where an action is clearly prohibited (haram), tools such as *muwazanat* do not apply. It is also not intended to be applied independently: Nonspecialists should not derive conclusions from it in isolation, nor should specialists employ it without broader scholarly context.

The framework of *fiqh al-muwazanat* for AI ethics

Concerns such as authority, responsibility, and identity show that the ethical challenges of AI are rarely simple. Benefits and harms are often inseparable and unevenly distributed. Addressing them piecemeal risks overlooking the trade-offs that shape real outcomes. Islamic law already offers a principled method for such situations: *fiqh al-muwazanat*, a methodological framework within Islamic jurisprudence for weighing competing considerations and evaluating consequences when multiple interests are at stake.[18] *Fiqh al-muwazanat* represents a systematic approach to resolving tensions between competing interests (*masalih*) and harms (*mafasid*) that is particularly valuable when addressing novel technological

[18] For a synthesis of these conditions and additional detail, see Suleiman Hani, "The Jurisprudence of Balancing (Fiqh al-Muwāzanāt): Principles and Applications," paper presented at the Assembly of Muslim Jurists of America 21st Annual Imams' Conference, Dallas, TX, August 22–24, 2025.

challenges not addressed in classical texts.[19] As articulated by Ibn al-Qayyim, "the foundation of the Sharia is wisdom and the safeguarding of people's interests in this world and the next. It is justice, mercy, benefit, and wisdom in its entirety."[20] This framework becomes essential when evaluating AI applications where benefits and harms coexist within the same technological capability.

Typologies of balancing (*muwazana*) in AI ethics

The *fiqh al-muwazanat* framework recognizes three distinct types of balancing that scholars must navigate when evaluating AI applications. Understanding these typologies is essential because AI technologies rarely present simple choices between clear good and clear harm. Instead, they typically involve complex scenarios where multiple benefits compete for priority, unavoidable harms must be minimized, or, most commonly, benefits and harms are inextricably intertwined within the same technology.

1. Balancing benefits (*muwazanat al-masalih*)

This type of balancing occurs when multiple beneficial AI applications compete for limited resources, attention, or implementation priority. For instance, a Muslim educational institution might need to choose between investing in AI-powered personalized learning systems that could dramatically improve student outcomes, or AI-enhanced Arabic language preservation tools that could help maintain Islamic cultural heritage. Both serve legitimate Islamic objectives, education (*taʿlim*) and preservation of religious knowledge (*hifz al-din*), but limited budgets may prevent pursuing both simultaneously.

The Islamic legal tradition provides clear principles for such situations. The obligatory (*wajib*) takes precedence over the recommended (*mandub*), communal obligations (*fard kifaya*) generally outweigh individual ones (*fard ʿayn*) when resources are limited, and benefits affecting larger populations take priority over

[19] Yūsuf al-Qaraḍāwī, *Fī fiqh al-awlawiyyāt: Dirāsa jadīda fī ḍaw' al-Qurʾān wa-l-Sunna* (Maktabat Wahba, 1996), 41.
[20] Ibn al-Qayyim al-Jawziyya, *I'lām al-muwaqqīʿīn ʿan Rabb al-ʿĀlamīn*, ed. Muḥammad ʿAbd al-Salām Ibrāhīm (Dār al-Kutub al-ʿIlmiyya, 1991), 3:11.

those benefiting fewer people.[21] When applying these principles to AI, scholars must consider factors such as: Which technology serves more fundamental objectives (*maqasid*)? Which addresses more urgent community needs? Which has greater potential for long-term positive impact?

2. Balancing harms (*muwazanat al-mafasid*)

Sometimes circumstances force a choice between two harmful outcomes, where avoiding both is impossible. In such cases, Islamic law requires choosing the lesser evil to minimize overall harm. The Prophet ﷺ exemplified this principle when he advised against rebuilding the Kaaba on its original foundations, despite this being religiously preferable, because it might harm the faith of newly converted Muslims.[22]

In the AI context, consider facial recognition technology. Law enforcement agencies argue it helps locate missing children and prevent terrorism, clear benefits for preserving life (*hifz al-nafs*). However, the same technology enables mass surveillance that is used to persecute Muslim minorities, as documented in China's treatment of Uyghurs.[23] If a Muslim-majority country must choose between completely banning such technology (potentially hampering legitimate security needs) or allowing it with the risk of abuse, scholars must determine which option minimizes overall harm to both individual privacy and collective security.

3. Balancing benefits and harms (*muwazanat al-masalih wa-l-mafasid*)

This most complex type of balancing addresses situations where the same AI application simultaneously produces benefits and harms that cannot be separated. The Qur'anic treatment of alcohol and gambling provides the classical model: "They ask you about wine and gambling. Say, 'In them there is great sin and [some] benefit for people. But their sin is greater than their benefit.'"[24] This verse establishes that Islamic law can weigh qualitatively different considerations,

---

[21] Ibn ʿAbd al-Salām, *Qawāʿid al-aḥkām*.
[22] *Ṣaḥīḥ al-Bukhārī*, no. 1586; *Ṣaḥīḥ Muslim*, no. 1333.
[23] Astha Anand, "Repression of Uyghur Muslims and the Freedom of Religious Beliefs in China," *Journal of Social Inclusion Studies* 8, no. 1 (2022): 23–36, https://doi.org/10.1177/23944811221085680.
[24] Qur'an 2:219.

benefits versus harms, and reach definitive conclusions.

Most AI applications fall into this category. Social media algorithms, for instance, help Muslims connect with scholars globally, access Islamic knowledge, and build community networks—significant benefits for religious practice. Simultaneously, these same algorithms can promote addiction, spread misinformation about Islam, and expose users to harmful content. The benefits and harms are produced by the same underlying technology and cannot be easily separated.

Essential conditions for applied *muwazanat* analysis

For scholars and institutions to effectively apply *fiqh al-muwazanat* to AI ethics, five fundamental conditions must be satisfied. Each of these conditions carries established pedigree in classical *usul al-fiqh*. Their consolidation into the five conditions of *muwazanat*, however, is a contemporary systematization rather than a fixed classical taxonomy. Together, they represent the essential foundations upon which the discipline is built. Neglecting them renders effective engagement in interest-balancing extraordinarily difficult,[25] and understanding each condition is crucial for those seeking to apply this framework to contemporary technological challenges.

1. Evaluating the objectives of Islamic law (*maqasid al-shariʿa*) and legal maxims

The *maqasid al-shariʿa* refers to the higher objectives and purposes that Islamic law seeks to achieve for humanity. Classical scholars identified five essential objectives that the Sharia aims to preserve: religion (*din*), life (*nafs*), intellect (*ʿaql*), lineage (*nasl*), and property (*mal*).[26] These are not merely abstract concepts but represent fundamental human interests that must be protected for both individual and societal well-being.

Understanding these objectives provides the necessary evaluative framework for

[25] For one such synthesis sample use cases, see Suleiman Hani, *The Fiqh of Muwazanat: The Jurisprudence of Weighing Benefits and Harms* (Waymark Publications, 2026), 15–30.
[26] Al-Shāṭibī, *al-Muwāfaqāt fī uṣūl al-sharīʿa*, 2:20.

assessing AI applications. For instance, when evaluating an AI medical diagnostic system, we consider how it serves the preservation of life (*hifz al-nafs*). When assessing AI-generated educational content, we examine its impact on preserving and developing human intellect (*hifz al-ʿaql*). This hierarchical framework helps scholars determine which benefits to prioritize and which harms to prevent.

Additionally, scholars must master key legal maxims (*qawaʿid fiqhiyya*) that guide decision-making. The principle that "preventing harm takes precedence over securing benefits" (*darʾ al-mafasid muqaddam ʿala jalb al-masalih*) proves particularly relevant for AI ethics. This means that if an AI application offers benefits but also poses significant risks of harm, Islamic law prioritizes preventing the harm even if it means forgoing the benefits. For example, if an AI system could improve educational outcomes but also risks exposing children to harmful content, the priority would be preventing the harm.

2. Verifying the effective cause (*tahqiq al-manat*)

This involves verifying whether contemporary AI phenomena genuinely fall under classical legal categories. As Ibn Taymiyya emphasized, this process is "the essence of *ijtihad*," requiring careful analysis of whether AI systems constitute mere tools (*alat*) or possess characteristics requiring novel jurisprudential treatment.[27] The process also involves examining the essential characteristics (*manat*) that determine how Islamic law treats something. For AI systems, relevant characteristics might include the degree of autonomy in decision-making, the predictability of outputs, the presence of learning capabilities that change behavior over time, and the potential for outputs that the programmer neither intended nor could foresee. This careful analysis prevents the misapplication of classical rulings to fundamentally new phenomena.

3. Considering context and reality (*fiqh al-waqiʿ*)

Fiqh al-waqiʿ literally means "jurisprudence of reality" and refers to the necessity

[27] Ibn Taymiyya, *Majmūʿ al-fatāwā*, ed. ʿAbd al-Raḥmān ibn Qāsim (Dār al-Wafāʾ, 2005), 19:203; Abū Isḥāq al-Shāṭibī, *al-Muwāfaqāt fī uṣūl al-sharīʿa* (Dār al-Maʿrifa, 1997), 4:89–90.

of understanding the actual circumstances, capabilities, and impacts of what we're evaluating. Ibn al-Qayyim warned that "whoever does not understand the reality [*al-waqiʿ*] of creation and the obligation [*al-wajib*] in religion will not understand God's rulings concerning His servants."[28] In the context of AI, this means scholars cannot issue meaningful guidance without understanding how AI actually works, its current capabilities and limitations, and its real-world impacts on individuals and society.

This condition requires engagement with empirical research and technical knowledge. For example, understanding that large language models like ChatGPT generate responses through statistical pattern matching rather than true comprehension is crucial for evaluating their use in religious contexts. Similarly, awareness of documented biases in AI training data directly impacts assessments of fairness and justice in AI applications.

Fiqh al-waqiʿ also demands understanding the social and economic contexts in which AI operates. How does AI adoption affect employment in Muslim communities? What are the psychological impacts of AI companions on human relationships? These contextual factors significantly influence the Islamic legal assessment of AI technologies. Ultimately, this necessitates ongoing engagement with current AI research and empirical studies on its effects.

4. Considering consequences (*iʿtibar al-maʾalat*)

Iʿtibar al-maʾalat means carefully considering the probable outcomes and long-term consequences of actions or rulings. Al-Shatibi established that "considering the consequences of actions is an established and intended principle in the Sharia."[29] This principle recognizes that an action's permissibility in Islamic law often depends not just on the action itself but also on what it leads to.

In AI ethics, this means looking beyond immediate benefits or harms to consider

[28] Ibn al-Qayyim al-Jawziyya, *Iʿlām al-muwaqqiʿīn*, 3:3; Shihāb al-Dīn al-Qarāfī, *al-Iḥkām fī tamyīz al-fatāwā ʿan al-aḥkām wa-taṣarrufāt al-qāḍī wa-l-imām*, ed. ʿAbd al-Fattāḥ Abū Ghudda, 2nd ed. (Dār al-Bashāʾir al-Islāmiyya, 1995), 112.
[29] Al-Shāṭibī, *al-Muwāfaqāt fī uṣūl al-sharīʿa*, 4:194, 198–200.

long-term impacts. For instance, while AI tutoring systems might provide immediate educational benefits, scholars must also consider potential long-term consequences: Will widespread adoption reduce human teacher employment? Could it diminish the mentor-student relationship that transmits not just knowledge but wisdom and character? Might children become overly dependent on AI, weakening their independent thinking abilities?

This forward-looking analysis includes the principle of *sadd al-dharaʾiʿ* (blocking the means to harm). If allowing certain AI applications could open pathways to greater harms, even if those harms are not immediate or certain, Islamic law may restrict them preventively. For example, even if current AI cannot independently issue religious rulings, allowing AI-generated religious content might normalize receiving spiritual guidance from machines, potentially leading to greater harm in the future.

5. Establishing the jurisprudence of priorities (*fiqh al-awlawiyyat*)

Fiqh al-awlawiyyat is the discipline of determining what takes precedence when multiple obligations, benefits, or harms compete for attention or resources. Determining which AI applications deserve immediate attention versus those that can be deferred requires systematic prioritization based on scope of impact, urgency, and alignment with fundamental Islamic objectives.[30] In the AI context, prioritization means determining which AI applications deserve immediate attention versus those that can be deferred. Should Muslim communities prioritize addressing AI in health care, education, finance, or surveillance? When resources are limited, which AI-related harms should be addressed first? This prioritization must be based on several factors:

- Scope of impact: Issues affecting entire communities take precedence over those affecting individuals.
- Severity of consequences: Life-threatening applications require more urgent attention than convenience features.
- Certainty vs. probability: Definite harms take precedence over speculative

[30] Al-Qaraḍāwī, *Fī fiqh al-awlawiyyāt*, 76–89.

benefits.

- Availability of alternatives: If non-AI alternatives exist, the urgency of AI adoption decreases.

Understanding these priorities prevents the misallocation of community resources and scholarly attention, ensuring that the most critical issues receive appropriate focus. An effective way for seminaries and research institutions to operationalize these five conditions would be to produce a living docket of AI issues, prioritized by *maqasid* and *awlawiyyat*, with periodic public updates.

Case study: AI and the question of fatwas

The value of *fiqh al-muwazanat* lies in application. Principles remain abstract until tested against real dilemmas where benefits and harms collide. Case studies allow us to see how Islamic jurisprudence engages AI not with ad hoc answers but through a disciplined framework.

With the typologies and conditions clearly established, we can now rigorously apply the jurisprudence of balancing to one of the most urgent contemporary dilemmas: the use of AI in issuing Islamic legal rulings (*ifta'*). This topic touches the very core of Islamic religious authority, accessibility, and trust, and raises fundamental questions about the nature of guidance in the digital age. While AI can assist by compiling prior rulings or identifying relevant texts, the possibility of automated or semiautomated fatwas introduces risks of decontextualization, misapplication, and erosion of scholarly discretion. Questions of intent (*niyya*), qualifications, and authority, central to *ifta'*, do not translate easily into computational logic.

To rigorously assess these competing benefits and harms, we now analyze the core typologies and conditions of *fiqh al-muwazanat*, examining each in turn.

Identifying the benefits (*masalih*)

Recent advances in AI technology, particularly large language models, have created systems capable of processing vast amounts of text and generating

human-like responses. Proponents of AI in Islamic legal consultation point to several potential benefits:

1. Unprecedented accessibility and dynamic personalization

The global Muslim population of approximately 1.9 billion faces a significant shortage of qualified Islamic scholars. This shortage is particularly acute in Muslim-minority regions, where access to qualified religious guidance may be extremely limited; for many in these demographics, the practical alternative to AI-generated guidance is not a local scholar, but a complete absence of guidance.

In this context, generative AI offers more than just the sheer reach of traditional online fatwa banks. It introduces the capability for dynamic personalization. Unlike static books or prerecorded databases that require users to find an exact match for their query, AI systems can synthesize highly tailored responses that take into account a questioner's specific variables, such as their designated *madhhab* (school of thought), geographic circumstances, and baseline level of Islamic knowledge. By dynamically adjusting the complexity and contextual framing of its outputs, AI can simulate the individualized attention typically reserved for one-on-one scholarly consultations, offering an interactive and highly responsive educational experience.

2. Enhanced research capabilities

Modern Islamic legal research requires consulting numerous classical texts, contemporary fatwas, and scholarly commentaries. AI systems can rapidly search through vast databases of Islamic legal literature, potentially identifying relevant precedents that human researchers might miss. While studies specific to Islamic legal AI are still emerging, research on general legal AI systems provides relevant insights. While early studies on legal AI noted significant limitations in accuracy,[31] current evaluations demonstrate that retrieval-augmented generation (RAG) systems, which combine advanced language models with specialized databases, have fundamentally altered the research landscape. By anchoring outputs strictly to

[31] Varun Magesh et al., "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," Stanford HAI, May 30, 2024.

verified corpora, these architectures heavily mitigate the risk of fabricated references and show immense promise in legal research contexts, though they still require scholarly oversight.[32]

3. Consistency in established matters

For issues where scholarly consensus (*ijmaʿ*) exists, AI could provide consistent responses that align with well-established rulings. This might reduce confusion arising from contradictory opinions on settled matters, particularly for Muslims without the scholarly background to distinguish between mainstream and eccentric positions.

Identifying the harms (*mafasid*)

The potential harms of AI-generated fatwas require extremely careful consideration, as they strike at fundamental aspects of Islamic religious practice:

1. Fundamental misunderstanding of legal context

Islamic legal reasoning (*ijtihad*) requires what al-Shatibi termed "*tahqiq al-manat*"—the careful verification of how general principles apply to specific circumstances.[33] This process demands not just textual knowledge but deep understanding of social contexts, individual circumstances, and the interplay between various legal considerations.

Early 2024 evaluations of legal AI tools demonstrated raw hallucination rates between 17 and 34 percent.[34] While frontier models (such as Claude Opus 4.6 and Gemini 3.1 Pro) have since made qualitative leaps in factual reliability, recent benchmarks highlight a persistent brittleness regarding deep contextual interpretation. While modern systems rarely fabricate citations, they still frequently misunderstand case contexts, cite irrelevant precedents, and fail to recognize when

[32] David Soong et al., "Improving Accuracy of GPT-3/4 Results on Biomedical Data Using a Retrieval-Augmented Language Model," *PLOS Digital Health* 3, no. 8 (2024): e0000568.
[33] Al-Shāṭibī, *al-Muwāfaqāt fī uṣūl al-sharīʿa*, 4:89.
[34] Magesh, "Hallucination-Free?" The study found LexisNexis's Lexis+ AI hallucinated 17 percent of the time, Thomson Reuters's Ask Practical Law AI hallucinated 17 percent of the time, and Westlaw AI-Assisted Research hallucinated 34 percent of the time.

underlying principles render a historical precedent inapplicable to a contemporary reality. This lack of holistic comprehension warrants significant caution, as the errors produced are no longer obvious fabrications, but rather highly plausible misapplications of the law.[35]

2. Absence of spiritual insight and wisdom

The role of a mufti extends far beyond mechanical application of legal rules. Ibn al-Qayyim emphasized that issuing fatwas requires understanding "the questioner's state, his needs, and what leads to his welfare."[36] This involves spiritual insight (*basira*), wisdom (*hikma*), and the ability to perceive underlying motivations and circumstances that may not be explicitly stated.

Current AI systems, despite their sophisticated pattern-matching capabilities, lack consciousness, spiritual understanding, and genuine comprehension. They cannot perceive the spiritual state of a questioner, assess their sincerity, or determine whether a stricter or more lenient approach would better serve their religious development.

3. Systematic biases in Islamic understanding

Early research into general-purpose large language models revealed stark, overt prejudices. For example, a 2021 Stanford study found that GPT-3 exhibited "severe bias" against Muslims, with violent associations appearing in two-thirds of responses when prompted with Muslim-related content.[37] Another analysis demonstrated that ChatGPT's responses on Islamic historical questions varied dramatically based on how questions were framed, producing contradictory narratives that reflected the biases in its training data rather than balanced scholarly perspectives.[38]

---

[35] Neel Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models," *Advances in Neural Information Processing Systems* 36 (2024).
[36] Ibn al-Qayyim al-Jawziyya, *Iʿlām al-muwaqqīʿīn*, 4:157.
[37] James Zou and Abubakar Abid, "Rooting Out Anti-Muslim Bias in Popular Language Model GPT-3," Stanford HAI, 2021.
[38] "ChatGPT on the 'Decline' of the Muslim World: A Case Study on the Biases of AI," *The Gazelle*, 2023.

Modern, purpose-built Islamic AI can moderate this concern using curated, scholar-approved databases. While this heavily mitigates the overt stereotypes, the structural risk profile of AI does not disappear as a result. Underlying foundational models are trained predominantly on high-resource languages (primarily English) and Western-centric datasets, which often embed deeply secular reasoning pathways. This shifts the risk from overt hostility to subtle ideological drift, where Islamic jurisprudence is subtly reframed to align with liberal, secular paradigms embedded in a particular AI's architecture.

4. Erosion of scholarly authority and methodology

The Islamic scholarly tradition maintains quality through mechanisms of peer review, scholarly credentials (*ijaza*), and the requirement for extensive study under qualified teachers. AI-generated fatwas bypass these crucial safeguards. The democratization of fatwa generation could lead to a situation where anyone can prompt an AI for religious rulings, potentially receiving responses that appear authoritative but lack proper scholarly grounding.

Essential conditions

Condition 1: Evaluating objectives of Islamic law (*maqasid al-shariʿa*) and legal maxims

The preservation of religion (*hifz al-din*) stands as the paramount objective of Islamic law. Accurate religious guidance serves this objective directly, while religious misguidance threatens it at the most fundamental level. When we examine AI-generated fatwas through this lens, we must ask: Does this technology primarily serve to preserve and strengthen proper religious understanding, or does it risk corrupting it? Is it generating new information (*iftaʾ*) or merely referencing existing information?

The accessibility benefits of AI, while significant, primarily serve to make religious information more accessible (*tahsiniyyat*), an improvement rather than a necessity (*daruriyyat*). While delayed access to a fatwa may cause practical

inconvenience, it rarely poses an existential threat to a community's faith; in contrast, the dissemination of erroneous religious guidance actively undermines the foundation of the religion. This hierarchical analysis suggests that protecting the integrity of religious guidance takes clear precedence over improving its accessibility.

Condition 2: Verifying the effective cause (*tahqiq al-manat*)

Applying *tahqiq al-manat* requires determining whether AI-generated religious content falls under the category of *ifta'* (issuing legal rulings) or merely *naql* (transmitting established rulings). This distinction is crucial because Islamic law treats these categories very differently.

Classical jurisprudence clearly distinguishes between creative legal reasoning (*ijtihad*), which requires qualified human judgment, and the mere transmission of established rulings, which even a child can perform by reading from a book. AI systems, lacking genuine comprehension, consciousness, and spiritual insight, cannot engage in actual *ijtihad*. At best, they might serve as sophisticated tools for *naql*, but even this limited role requires careful constraints to prevent the appearance of independent legal reasoning.

Condition 3: Considering the current reality (*fiqh al-waqi'*)

Evaluating the current reality (*fiqh al-waqi'*) requires acknowledging the rapid evolution of AI capabilities in religious and legal contexts. While early systems suffered from high hallucination rates, contemporary foundational models have achieved exceptional baseline accuracies on complex reasoning tasks. However, this high performance introduces a new challenge: The technology's rapid deployment has outpaced quality control mechanisms.

Unlike traditional Islamic scholarship, which develops through years of careful study and peer review, AI systems can be updated overnight, potentially introducing new biases or errors without users' knowledge. The "black box" nature of large language models means that even their creators cannot fully explain why they generate specific responses. Built from billions of parameters that interact in

highly complex ways across many layers, AI responses are the result of statistical patterns, not a step-by-step, human-readable chain of reasoning. We see what goes in and what comes out, but we cannot clearly trace or explain the exact internal reasoning that produced that answer.

Condition 4: Considering consequences (*iʿtibar al-maʾalat*)

The probable and long-term implications of normalizing AI-generated religious guidance extend far beyond immediate concerns about accuracy. Consider the potential trajectories:

1. Erosion of human spiritual guidance: As people become accustomed to receiving instant religious answers from AI, the tradition of seeking guidance from living scholars, with all the spiritual benefits such relationships provide, may gradually disappear. Earlier, less capable information technologies have disrupted traditional Islamic authority and the scholar-seeker relationship, from the advent of the printing press to the rise of the internet, social media, and fatwa banks.[39]
2. Theological drift: AI systems trained on internet data inevitably absorb and reproduce contemporary biases. Over time, this could lead to a gradual shift in religious understanding, moving away from scholarly interpretations grounded in tradition toward whatever perspectives dominate online discourse. While most AI usage today is prone to this consequence, a purpose-built AI with carefully constructed anchoring can mitigate this concern.
3. Loss of contextual jurisprudence: Islamic law's strength lies partly in its ability to provide context-specific guidance. AI systems, operating through pattern matching rather than true understanding, may push Islamic legal thought toward rigid, decontextualized applications.

[39] Metin M. Coşgel, Thomas J. Miceli, and Jared Rubin, "The Political Economy of Mass Printing: Legitimacy and Technological Change in the Ottoman Empire," *Journal of Comparative Economics* 40 (2012): 357–71, https://doi.org/10.1016/j.jce.2012.01.002; Amamur Rohman Hamdani, "Fatwa in the Digital Age: Online Muftī, Social Media, and Alternative Religious Authority," *Hikmatuna: Journal for Integrative Islamic Studies* 9 (2023): 53–63, https://doi.org/10.28918/hikmatuna.v9i1.966.

Following the principle of *sadd al-dharaʾiʿ* (blocking paths to harm), even if current AI systems could be constrained to avoid major errors, permitting their use for religious guidance opens pathways to these greater future harms.

Condition 5: Establishing the jurisprudence of priorities (*fiqh al-awlawiyyat*)

Given the fundamental importance of preserving authentic religious guidance, and recognizing that alternative solutions exist (such as using AI to assist human scholars rather than replace them), the priority calculation becomes clear. The potential harm to religious integrity far outweighs the convenience benefits of AI-generated fatwas.

Recommended guidelines for issuing Islamic legal rulings (*iftaʾ*) based on *muwazanat* analysis

Based on this comprehensive analysis, the following guidelines emerge:

1. Prohibition on autonomous fatwa generation

While AI may be utilized to retrieve and transmit basic, settled knowledge (*naql*), it must not be used to independently issue fatwas, synthesize novel legal reasoning (*ijtihad*), or apply general rules to specific personal circumstances. In these domains of active legal judgment, the harms to religious authority, accuracy, and spiritual guidance decisively outweigh any benefits of increased accessibility.

2. Permitted use as research assistance

AI may serve as a research tool for qualified scholars, helping to search through texts, compile relevant precedents, and organize information. However, all analysis, reasoning, and conclusions must come from human scholars.

3. Educational applications with clear limitations

AI can assist in teaching established rulings and basic Islamic principles, provided clear, prominent disclaimers indicate that this is educational material, not religious

guidance. Educational material involves the transmission (*naql* and *taʿlim*) of established, general knowledge, whereas religious guidance (*iftaʾ*) involves the application of rulings to specific, personal circumstances. Furthermore, the system used should explicitly direct users to consult qualified scholars for personal religious questions. Additionally, the content must be regularly reviewed by qualified scholars for accuracy.

4. Institutional oversight requirements

Muslim organizations implementing any AI tools in religious contexts should establish committees of qualified scholars and technical experts to: (1) evaluate tools before implementation, (2) monitor ongoing performance and identify biases, (3) ensure compliance with Islamic legal principles, and (4) review and approve any updates or modifications.

5. Transparency and accountability

Any AI system used in Islamic educational contexts must maintain transparency about: (1) its limitations and lack of spiritual understanding, (2) the sources of its training data, (3) the process for reporting and correcting errors, and (4) the human scholars responsible for oversight.

This analysis demonstrates how the systematic application of *fiqh al-muwazanat* provides clear, principled guidance for addressing novel technological challenges while maintaining fidelity to Islamic legal methodology and objectives.

Additional problems in need of Islamic research and *ijtihad*

With the imminent growth of AI, a focused research agenda that addresses urgent ethical and theological questions across social, economic, and religious spheres is needed. These questions are inherently interdisciplinary, requiring Islamic scholars to engage alongside experts in fields such as computer science, law, and philosophy. Here, we introduce an inexhaustive set of pressing areas that stand out

as needing deeper research and sustained engagement. Other questions will continue to emerge as AI develops, yet even this small set illustrates how urgently Islamic scholarship must grapple with the challenges ahead. We provide brief primers on the core issues and anchoring principles around these questions, and call upon multidisciplinary teams to undertake deeper scholarly treatments of them.

AI, epistemology, and the pursuit of truth

One of the most pressing concerns regarding AI lies in the domain of authority and knowledge. This area sits at the very center of the preservation of religion (*hifz al-din*) and the epistemic trust upon which Muslim communities rely. AI systems that generate answers to religious questions, draft sermons, or summarize rulings blur the line between basic research assistance and Islamic legal pronouncement (*ifta'*). Their outputs often carry an aura of absolute certainty and neutrality, even though they are merely statistical predictions shaped by their training data. Critics of this view often argue that human scholars are similarly "programmed" by their teachers and cultural context. In its strongest form, this objection rests on a reductive materialism that the Islamic tradition does not concede; a human scholar's judgment is anchored in moral agency, spiritual insight, and accountability before Allah, not merely in the weights of prior exposure. Yet the distinction does not depend solely on that metaphysical dispute. Even granting the analogy at the level of training, the scholar and the model differ functionally in ways decisive for *ifta'*: The scholar forms an intention (*niyyah*), can articulate and be held to the chain of reasoning by which a ruling was reached, and bears personal liability for it before both God and the community. In contrast, the model yields outputs through opaque statistical processes for which no accountable author can supply, or answer for, the underlying reasoning. What the category of *ijtihad* presupposes, then, is this accountable authorship—not a theory of what cognition is made of.

The reliance on statistical predictions creates serious risks in the context of *ifta'*: the generation of decontextualized rulings, the misapplication of sacred texts, and the circulation of material that appears deeply authoritative but lacks any genuine scholarly grounding. A human scholar can be held accountable to explain the

precise jurisprudential methodology that led to a ruling, whereas the underlying reasoning of a neural network is largely opaque. While the ability of these tools to instantly sift through vast textual corpora makes them undeniably attractive, we must ask: How can the integrity of scholarly authority be protected in an environment where machine-generated content is so ubiquitous?

To answer this, we must first recognize how Islam understands truth. In the Islamic worldview, truth is not merely a property of sentences; it is an undeniable reality grounded in the Divine. Al-Haqq (the Truth) is one of the names of God, and the Qur'an repeatedly states that the heavens and the earth were created *bi-l-haqq* (in truth).[40] Because reality itself is anchored in the Divine, our pursuit of knowledge is a sacred and moral duty, not just an intellectual exercise.

Large language models (LLMs), however, do not operate within this moral universe. They are not truth-tracking agents; they are extraordinarily powerful next-word predictors. Their underlying architecture is designed to optimize for plausibility, not truth. They do not "know" what they are saying. Consequently, AI systems are infamous for "hallucinations": confidently inventing quotes, fabricating historical events, or even generating fake Qur'anic verses and hadith.

Because AI routinely produces both accurate summaries and complete fabrications with the exact same tone of unwavering confidence, it preys upon so-called "automation bias." Studies on generative AI show that its sheer fluency frequently overrides our critical faculties, making users deeply reticent to actually fact-check its claims.[41] AI-generated content in the public sphere cannot be granted the presumption of plausibility. Instead, it must be evaluated at the most basic epistemic baseline: between truth (*sidq*) and falsehood (*kadhib*). The Qur'anic demand for evidence is explicit: "Say, 'Produce your proof, if you should be truthful.'"[42] This Prophetic principle dictates that the burden of proof lies upon the

[40] Qur'an 15:85; 21:18.
[41] S. Shyam Sundar and Jinyoung Kim, "Machine Heuristic: When We Trust Computers More Than Humans with Our Personal Information," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2019), 1–9, https://doi.org/10.1145/3290605.3300768.
[42] Qur'an 2:111, 27:64.

claimant. It is not the responsibility of the Muslim audience to assume a machine's output is true; it is the burden of the person utilizing the output to verify it.

Navigating this reality does not require us to invent a new ethical framework; rather, it requires the revival of a classical Islamic epistemic ethic: *naqd* (rigorous critical evaluation). Historically, when the early Muslim community faced a proliferation of fabricated reports and malicious liars (*kadhdhabin*)—particularly in diverse, intellectually active centers like Kufa—scholars did not respond with a framework of "calibrated trust," a posture that assumes a baseline of reliability that is merely adjusted when errors occur. They responded with disciplined suspicion: an epistemological baseline of doubt where the burden of proof rests entirely upon the transmitter. They developed the rigorous sciences of hadith criticism and biographical evaluation (*ʿilm al-rijal*) because the preservation of the religion demanded skepticism of unverified claims.[43] The rigorous development of the science of narrators (*ʿilm al-rijal*) served as an independent epistemological infrastructure to protect the early Muslim community from intellectual dependency and the fabrication of knowledge. Relying entirely on foundational AI models trained on overwhelmingly Western datasets with embedded biases is akin to accepting a narrative without verifying its chain of transmission.

This culture of disciplined suspicion is exactly what is required today in an environment saturated with AI-generated content. For Muslim scholars and the general public alike, the correct posture toward AI is neither naïve adoption nor categorical rejection, but rigorous verification (*tabayyun*).[44] By treating AI not as an authority, but as a sophisticated yet profoundly fallible tool, we can utilize its efficiencies while guarding our minds and our faith with the same vigilance championed by the scholars of our past. Discerning *al-haqq* entails calibrated skepticism in AI outputs, disciplined reason, and the policing of error through verification. Recent work in Islamic AI ethics argues for precisely this pluralist benchmarking: preserving metaphysical commitments about *al-haqq* while

[43] For a detailed discussion on the rise of systematic hadith criticism (*naqd*) in response to the proliferation of fabricated reports, see Jonathan A. C. Brown, *Hadith: Muhammad's Legacy in the Medieval and Modern World* (Oneworld Publications, 2009), 67–75. See also Scott C. Lucas, *Constructive Critics, Ḥadīth Literature, and the Articulation of Sunnī Islam* (Brill, 2004), 115–22.
[44] Qur'an 49:6.

operationalizing legal-ethical tools (e.g., *qiyas*, *istislah*, *sadd al-dharaʾiʿ*) to decide when probabilistic assistance serves or subverts the just ends of Sharia.[45] Such a framework honors the Islamic ontology of truth while acknowledging the distinctive, distribution-fit epistemics of AI.

Liability and ownership

Islamic teachings consistently emphasize that every individual will be held accountable before Allah for their actions. The Qur'an declares, "Every soul is held in pledge for what it has earned,"[46] and the Prophet ﷺ said, "Each of you is a shepherd, and each of you is responsible for his flock."[47] In an age of rapidly evolving artificial intelligence, these ethical and legal responsibilities take on renewed urgency. Muslim scholars and technologists alike are compelled to ask: What is my responsibility before Allah in using or developing AI? And how can the foundational sources of Islam help assess the outcomes of actions mediated by AI systems? One established framework for engaging these questions is found in the Islamic laws of liability (*daman*), which distinguish between spiritual accountability before Allah and legal liability for material harm. Classical jurists outlined various scenarios in which a person may be (1) legally liable but not sinful, (2) sinful but not legally liable, or (3) both liable and sinful.[48]

To apply these frameworks to artificial intelligence, we must first establish the exact nature of the technology. Because AI systems lack sentience, will, and moral agency, they are fundamentally nothing more than programmable artifacts and sophisticated tools. Consequently, an AI model itself can never be classified as a direct causer of harm or an indirect initiator in Islamic law, as both categories pertain exclusively to morally accountable human agents (*mukallaf*).

Instead, the classical Islamic legal tradition offers three main principles for

---

[45] Ezieddin Elmahjub, "Artificial Intelligence (AI) in Islamic Ethics: Towards Pluralist Ethical Benchmarking for AI," *Philosophy & Technology* 36 (2023): article 73.
[46] Qur'an 74:38.
[47] *Ṣaḥīḥ al-Bukhārī*, no. 893; *Ṣaḥīḥ Muslim*, no. 1829.
[48] Badr al-Dīn Muḥammad ibn Bahādur al-Zarkashī, *al-Manthūr fī al-qawāʿid* (Wizārat al-Awqāf wa-l-Shuʾūn al-Islāmiyya, 1985), 2:344; Zayn al-Dīn ibn Ibrāhīm Ibn Nujaym, *al-Ashbāh wa-l-naẓāʾir* (Dār al-Kutub al-ʿIlmiyya, 1999), 85.

assessing the liability of the human actors involved:

1. Direct causer (*mubashir*): Held legally responsible for harm caused directly by their actions, even if the act itself was not explicitly prohibited (haram).
2. Indirect initiator (*mutasabbib*): Liable if their action was impermissible, provided the connection to the harm is firmly established.
3. Combined actions: In cases involving both direct and indirect actors, liability typically rests with the direct actor unless the harm cannot rationally be traced to them.

When applied to AI systems, these principles are clear: They dictate that liability must be directed at the humans who design, deploy, and interact with the technology.

While generative AI models exhibit designed unpredictability, treating this trait as comparable to the instinctual agency of a living creature obscures the core moral issue. For example, the liability here is not analogous to that of a parent of a minor child who accidentally breaks an item at a store, or a trained dog that causes inadvertent harm. This is because the behavior of AI systems can be explicitly constrained through human design choices, such as implementing guardrails, certainty thresholds, verification layers, and human-in-the-loop approvals. Harms generally arise when developers or corporations choose to deploy immature or opaque systems into live, high-risk environments, often driven by market speed or profit, rather than from an inherent inability to constrain the system.

From an Islamic perspective, knowingly exposing people to foreseeable harm through such negligence is ethically unacceptable. Therefore, the burden of liability (*daman*) firmly remains on the human programmer, deployer, or user depending on the vector of caused harm. Islamic legal theory provides a rigorous framework for navigating these issues, ensuring that accountability is always traced back to human choices, foreseeability, negligence, and intent.

Intellectual and creative attribution

Classical Islamic ethics already provides robust frameworks for addressing misattribution,[49] deception,[50] and the unauthorized exploitation of another's labor.[51] In these domains, AI does not necessarily introduce novel moral dilemmas; instead, it drastically lowers the barrier to entry, obscures the mechanics of theft, and exponentially amplifies the scale of such potential ethical violations.

Moreover, responsibility and accountability for claims can be obfuscated due to novel error vectors in generative AI when performing background research and verification. In Islamic ethics, authorship is tied not only to producing words or images but to the intent, *ijtihad* (effort), and trustworthiness of the one who creates content. Passing off a khutbah drafted by ChatGPT as one's own scholarship, or a fatwa summary compiled by an algorithm as a jurist's considered ruling, can undermine the more direct connection between intention, effort, and ownership in classical research methods. Passing off machine-generated work as human erodes trust and blurs responsibility for the content being taught or shared; new paradigms of research and verification are needed in order to preserve trust in the integrity of religious voices.

The problem also extends to the data on which these systems are built. Vast corpora, including copyrighted works, sacred texts, and personal material, are often scraped without consent,[52] with their outputs reproducing these materials in altered or decontextualized forms. This raises questions of the protection of property (*hifz al-mal*) and fairness, since Islamic law generally prohibits benefiting from the labor or property of others without their permission.[53] Contemporary scholarship is

[49] Abū Zakariyyā Yaḥyā al-Nawawī, *al-Majmūʿ sharḥ al-muhadhdhab* (Dār al-Fikr, 1997), 1:93. Al-Nawawī quotes the earlier scholar Abū al-Ṭāhir al-Silafī, stating, "From the blessings of knowledge is to attribute the saying to its speaker." See also Jalāl al-Dīn al-Suyūṭī, *al-Fāriq bayna al-muṣannif wa-l-sāriq* (ʿĀlam al-Kutub, 1998), which explicitly addresses the ethics of authorship and plagiarism.

[50] *Ṣaḥīḥ Muslim*, no. 101. Regarding the concealment of reality (*tadlīs*), see Muḥammad ibn Aḥmad Ibn Rushd, *Bidāyat al-mujtahid wa-nihāyat al-muqtaṣid* (Dār al-Ḥadīth, 2004), 3:154–56.

[51] Muwaffaq al-Dīn ʿAbdullāh ibn Aḥmad Ibn Qudāma, *al-Mughnī* (Dār ʿĀlam al-Kutub, 1997), 7:360–62.

[52] Tom Gerken, "New York Times Sues Microsoft and OpenAI for 'Billions,'" *BBC*, December 27, 2023, https://www.bbc.com/news/technology-67826601.

[53] There are some nuances regarding, for example, religious knowledge or deceased authors. See Jamaal al-Deen Zarabozo, "The Copyright Issue," *MuslimMatters*, January 8, 2010, https://muslimmatters.org/2010/01/08/the-copyright-issue/; Irshaad Sedick, "Is Asking AI About Copyrighted

needed to extend classical paradigms of research and accountability to the era of generative AI, and to explore the broader societal implications of private companies profiting off of the collective output of training data without explicit consent of or compensation for their creators.

Environmental stewardship

Artificial intelligence is often imagined as an ethereal cloud of algorithms, but in practice, it is firmly anchored in vast physical infrastructure. The foundation of AI consists of graphics processing units (GPUs), tensor processing units (TPUs), specialized silicon chips, and sprawling data centers. This reality ties the advancement of AI directly to the global political economy of hardware extraction, mineral mining, and immense electricity consumption.[54]

Like much of modern industrial infrastructure, these facilities come with profound environmental costs. The modern internet already relies on energy-intensive systems that leave a significant ecological footprint, but the rise of generative AI magnifies these costs exponentially, demanding unprecedented resources at a global scale. AI represents a spectrum of resource intensity. At one end are huge foundation models, the training of which consumes extraordinary amounts of electricity and water. At the other end are everyday inference tasks that run quietly in the background of ordinary life. While foundational training is highly resource-intensive, the ecological cost of individual inference is actually decreasing rapidly. Recent empirical data from 2025 demonstrates that the energy required for a standard generative text prompt has dropped significantly, consuming merely 0.24 Wh (equivalent to watching television for less than nine seconds).[55] However, because inference is constant, distributed, and woven into our daily routines, the sheer exponential volume of these efficient queries, combined

---

Material Permissible in Islam?," SeekersGuidance, September 6, 2023, https://seekersguidance.org/answers/halal-and-haram/is-asking-ai-about-copyrighted-material-permissible-in-islam/.

[54] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021), 30–34.

[55] Cooper Elsworth, Keguo Huang, David Patterson, et al., "Measuring the Environmental Impact of Delivering AI at Google Scale," *arXiv*, August 21, 2025, https://arxiv.org/abs/2508.15734.

with the massive upfront infrastructural costs of model training, ensures that the aggregate global energy demand remains a profound ecological challenge.

From an Islamic paradigm, this reality demands a rigorous ethic of stewardship (*khilafa*). Because total disengagement from this technological spectrum is practically unrealistic, Muslims must carefully weigh the harms and benefits across these layers. Where the ecological costs outweigh the communal benefits, Islamic ethics allows for, and sometimes necessitates, abstention or restraint. For the Muslim professional, working within the system to push for smaller, more efficient models, demanding transparency in compute and data usage, and developing benchmarks for resource responsibility are all critical expressions of preserving the earth (*hifz al-biʾa*).

Digital sovereignty and infrastructure dependence

While the ecological footprint of AI is a highly consequential concern, the concentration of AI infrastructure and computational resources in a small number of corporations and states represents a more fundamental, metaethical challenge. This monopolization shapes exactly how AI is developed, deployed, and governed, and it acts as the structural root underlying many downstream harms. For the global Muslim *ummah*—which largely resides within developing countries and the Global South—this is not merely an issue of market competition; it is a question of profound civilizational significance and a shield against new forms of neocolonial dependency.

The training of foundation models and the provision of large-scale compute resources are overwhelmingly dominated by a handful of Western, primarily US-based, technology firms. The critical issue is not simply whether developing nations will have consumer access to AI, but whether they will possess the agency to shape its foundational infrastructure. Furthermore, because these foundational models are trained predominantly on high-resource languages (primarily English) and Western datasets, they suffer from severe language limitations and structural

biases.[56] These models systematically marginalize languages critical to the Muslim world, such as Arabic, Urdu, Farsi, and Malay, while embedding Western-centric, often secular assumptions into their very architecture.

Historically, Muslim scholars recognized that intellectual and spiritual sovereignty required robust, independent infrastructure. The rigorous development of hadith criticism, particularly the meticulous evaluation of transmission chains (*isnad*) and narrator reliability (*al-jarh wa-l-taʿdil*), was essentially a massive infrastructural project designed to protect the integrity of Islamic epistemology from foreign adulteration and fabrication.[57] Consequently, Muslim engagement with AI cannot rely on a single strategy of passive consumption or the mere mitigation of downstream harms. There is a pressing need to address this dependency at the root. Investing in digital sovereignty means developing shared, localized infrastructure in Muslim-majority societies and the broader Global South, such as regional cloud networks, indigenous compute clusters, and foundation models trained on datasets carefully curated to reflect and respect orthodox Islamic commitments. Without this, Muslim communities risk a profound geopolitical and epistemological dependence on the few entities that control the digital future. We believe this is an area ripe for entrepreneurship, innovation, and scholarship to help enable the promise of the generative AI era to extend to the entire world.

Recommendations for general Muslim engagement

The rapid spread of AI ensures that its impact will reach practically every Muslim household, classroom, workplace, and institution. Islamic ethics provides a framework for foresight and accountability, but it also requires translation into general practice across different layers of society. The following opportunities outline how individuals, communities, and professionals can contribute at a grassroots level to meaningfully shape AI's adoption in line with Islamic values.

[56] Shakir Mohamed, Marie-Therese Png, and William Isaac, "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence," *Philosophy & Technology* 33, no. 4 (2020): 659–84.
[57] Brown, *Hadith*, 70–75.

At the most immediate level of engagement, Muslims encounter AI as everyday users, and the priority in this context is building literacy and resilience. Individuals should cultivate habits of verification (*tabayyun*) when assessing AI-generated information, with the awareness that misinformation and deepfakes are often designed intentionally to manipulate. Families, and especially parents, bear the responsibility of modeling ethical technology use, setting boundaries for healthy reliance on AI and protecting children from exploitative or addictive applications. Attention to personal privacy, critical evaluation of AI outputs, and mindful use of AI tutoring or therapy tools provide a baseline of ethical literacy that can be established within Muslim homes.

Religious and community leaders shape how Muslims collectively perceive and use AI. Imams, teachers, and local leaders can guide their communities toward balanced approaches that neither exaggerate the benefits of AI nor dismiss it entirely. Through sermons, curricula, and workshops, they can frame AI as a tool that demands oversight rather than as an authority to be trusted without question. Community institutions should set clear policies for their own use of AI, ensuring that mosque communications, educational resources, and khutbah preparation are reviewed by qualified humans and labeled appropriately. Leaders should also adopt referral protocols for AI-related harms, including counseling pathways for those facing deepfake harassment or dependence on chatbots for companionship or therapy. By establishing norms early, they can prevent harmful practices from becoming ingrained in communal life.

Muslims working in professional and institutional settings have opportunities to influence how AI is built and governed. Technologists can advocate for responsible design, contribute to testing models for bias against Islam alongside broader systemic prejudices like racism and elitism, and help create datasets that reflect Islamic perspectives. Those in product development and policy roles have the ability to anticipate potential harms before systems are deployed, establish clear avenues for accountability when harms do occur, and uphold user privacy as a matter of trust. Institutions should collaborate to build shared infrastructure, such as benchmarking hubs and mentorship networks, that strengthen Muslim capacity in the AI era. They are also well-placed to explore strategies for digital

sovereignty, including independent datasets and infrastructure. In addition, Muslim professionals and community organizations can anticipate job displacement by supporting workforce retraining and reskilling, ensuring that vulnerable groups are not left behind.

Takeaways and future directions

Artificial intelligence is not a distant concern. It is an immediate ethical frontier, and the cost of hesitation is already clear. Our experience with social media demonstrated that delayed ethical engagement allows harmful structures to harden; with AI, the stakes are significantly higher. A reactive posture is no longer tenable. What is required is deliberate, principled intervention grounded in a coherent moral framework.

This study has argued that the Islamic intellectual tradition is not only relevant to this challenge but uniquely equipped to address it. By foregrounding accountability before both society and God, it offers a more comprehensive alternative to prevailing legalistic approaches that often fail to capture the full moral weight of technological harm. The application of *fiqh al-muwazanat*, guided by the *maqasid al-shariʿa*, provides a disciplined method for evaluating AI through its real-world consequences.

The case of AI-generated fatwas underscores the urgency of this approach. Increased accessibility cannot justify undermining the integrity of religious authority or risking distortion in matters of faith. Where harm is credible and far-reaching, restraint is a necessity and outweighs any speculative benefit. This principle extends beyond a single application: Any deployment of AI must be judged by whether it preserves or compromises essential human and religious goods.

The path forward demands coordinated effort. Key priorities include establishing a rigorous epistemological framework for AI outputs, treating them as provisional and subject to verification (*tabayyun*). It requires refining doctrines of liability (*daman*) to ensure that responsibility remains firmly with human agents. Questions

of algorithmic bias, data governance, and environmental cost must also be addressed as integral concerns as a necessity for safeguarding communal well-being. Most importantly, these insights must be institutionalized. The formation of a cross-disciplinary consortium of Muslim scholars, technologists, and legal experts is necessary. Such a body can set standards, audit systems, and advocate effectively in policy arenas. Without this level of organization, ethical guidance risks remaining theoretical while technological development accelerates unchecked.

AI is set to shape the conditions of human life in profound ways. The question is not whether to engage, but how and on whose terms. A principled, proactive approach offers the possibility of mitigating harm and directing technological development toward outcomes that genuinely serve the common good. By fostering digital sovereignty and investing in our own institutional capacity, we can actively shape a technological future that aligns with our values. This work represents a collective responsibility, an opportunity to ensure that these powerful new tools are harnessed not for profit or power, but for the genuine benefit of humanity.

Appendix: Demystifying generative AI[58]

To help contextualize the frameworks expounded in this paper, this appendix provides an accessible technical primer designed to demystify generative AI. By unpacking its key features, history of development, impact, and operational reality, we hope to equip readers with the foundations needed to begin assessing AI use.

The AI of today often appears mysterious, even magical, to those outside the field.[59] As with earlier technologies, this sense of mystery can create misplaced trust or

[58] This technical exposition focuses heavily on LLMs, given their rapid rise and profound impact. Nongenerative AI applications, including ranking, optimization, classification, and computer vision systems, can be prone to other ethical issues and different failure modes as compared to LLMs. We have restricted our scope to focus on generative AI.

[59] As an aside, Kate Crawford, a senior researcher in the field, quips that the current form of AI is neither artificial nor intelligent, with its heavy reliance on natural resources as well as output of human work. Zoë Corbyn, "Microsoft's Kate Crawford: 'AI Is Neither Artificial nor Intelligent,'" *The Guardian*, June 6, 2021,

exaggerated fear, both of which have unintended consequences. An example from recent history concerns the introduction of the radio, or wireless telegraphy more generally, in the Middle East in the early twentieth century. When the radio was first introduced, some influential religious scholars opposed it, fearing it was a tool of sorcery or instrument of the devil: They heard disembodied voices without grasping the underlying technology that enabled them. King Abdulaziz ibn Saud arranged for a pragmatic demonstration: He had reciters in Mecca read verses from the Qur'an over the radio, which the scholars heard clearly more than 500 miles (850 km) away in Riyadh. This helped successfully persuade them that the technology was a tool that could be used for good, and ultimately led to its official sanctioning and adoption.[60]

We argue that the products of artificial intelligence are analogous in that regard: Its outputs can feel uncanny or even humanlike, yet they are the result of definable systems and methods. As described within this paper, they are not entirely value neutral. To engage AI ethically and responsibly, we must first strip away this aura of mystery and understand what is really happening inside.

The technology most responsible for this perception is the rise of large language models (LLMs), systems designed to process and generate human language. For decades, the idea that computers could effectively understand and generate human language seemed closer to science fiction than reality. Early attempts at computational linguistics, which date back to the mid-twentieth century,[61] struggled with even simple grammatical sentences, leading to widespread skepticism. Later public-facing applications in the late 1990s and 2000s were isolated and modular, such as voice-recognition systems in customer service applications and recommender systems at Amazon and Netflix that personalize every viewing experience.

---

https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent.

[60] Robert Lacey, *The Kingdom: Arabia and the House of Sa'ud* (Harcourt Brace Jovanovich, 1981); Khayr al-Dīn al-Ziriklī, *Shibh al-jazīra fī ʿahd al-malik ʿAbd al-ʿAzīz*, 2nd ed. (Dār al-ʿIlm li-l-Malāyīn, 1985).

[61] "The First Public Demonstration of Machine Translation Occurs," History of Information, https://www.historyofinformation.com/detail.php?id=666.

However, over the last decade, rapid advancements in machine learning and computational capabilities have turned language modeling into one of AI's most astonishing success stories. Today, AI-driven language models, powered by predictive neural networks, produce coherent and contextually relevant stories, summarize documents, answer questions, translate documents seamlessly across languages, write code, and power intelligent assistants that engage in meaningful conversations. These applications represent a step function change. They are now embedded comprehensively in user experiences, deciding Uber driver matches, evaluating job candidates,[62] and flagging fraudulent transactions. The largest models are able to solve International Olympiad–level problems,[63] convincingly pass bar exams,[64] detect cancer with better accuracy than experts,[65] plan itineraries and recipes, and deploy sinister ends, such as identifying human targets in Gaza and Ukraine.[66]

Language models: History and progress

A language model is fundamentally a statistical tool designed to predict what word or phrase will come next, given some text.[67] The concept is straightforward, yet its implications are profound. Consider the following example. If given the sentence fragment, "Before praying, Ahmad performed _____," a language model, based on its prior exposure to extensive written language, assigns probabilities to various completions:

---

[62] For example, see Russell Reynolds Associates, https://www.russellreynolds.com/en/.
[63] Himanshi Lohchab, "OpenAI's o1 Takes a Leap with Model That Reasons Like Us," *The Economic Times*, September 16, 2024, https://economictimes.indiatimes.com/tech/technology/openais-o1-takes-a-leap-with-model-that-reason-like-us/articleshow/113373259.cms?from=mdr.
[64] Pablo Arredondo, "GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession," Stanford Law School, April 19, 2023, https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/.
[65] Radboud University Medical Center, "AI Better Detects Prostate Cancer on MRI Than Radiologists," *Science Daily*, June 12, 2024, https://www.sciencedaily.com/releases/2024/06/240612113341.htm
[66] Yasmeen Serhan, "How Israel Uses AI in Gaza—And What It Might Mean for the Future of Warfare," *Time*, December 18, 2024, https://time.com/7202584/gaza-ukraine-ai-warfare/.
[67] "Introduction to Large Language Models," Machine Learning course, Google Developer Program, https://developers.google.com/machine-learning/crash-course/llm.

- “wudu” (22.3%)
- “dhikr” (8.7%)
- “sunnah” (5.5%)

This predictive capability might appear trivial, but it is precisely this simplicity that underpins the model’s immense power. Because the entire internet, digital books, research articles, and transcripts of conversations can serve as training examples, every single piece of digitized text becomes a potential learning resource. The sheer scale of available training data allows these models to internalize vast amounts of knowledge about grammar, context, semantics, cultural references, and general world knowledge. Coupled with dramatic increases in computational power, particularly advances in specialized hardware like Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), this unprecedented combination of enormous datasets and intensive computation has enabled the creation of AI systems capable of performing an extraordinarily wide, versatile range of tasks. Thus, a language model learns far more than just predicting the next word: It learns the patterns, structures, and meanings deeply embedded in human language and thought.

Early computational linguistics relied on simple statistical methods like Markov chains, which predicted each word solely based on the preceding word. These approaches produced clunky, mechanical outputs and reinforced skepticism that computers could ever handle natural language effectively. Artificial neural networks, a powerful method of statistical machine learning, then emerged as the preferred approach. By autonomously discovering complex linguistic patterns in vast amounts of data rather than relying on rigid, hand-coded rules, these deep learning systems overcame earlier bottlenecks and quickly became central to everyday technology across research and industry. The field advanced in the early 2010s with deep learning methods, including recurrent neural networks and long short-term memory networks, which captured more nuanced linguistic context.

A decisive breakthrough came in 2017 with the introduction of the transformer architecture.[68] Unlike earlier models that processed text sequentially, transformers implemented a new mechanism called self-attention. To grasp the importance of this innovation, consider the following sentence: "The trophy doesn't fit in the suitcase because it is too large." In understanding this sentence, one needs to determine whether the pronoun "it" refers to "the trophy" or "the suitcase." A transformer's self-attention mechanism allows each word in the sentence to directly assess its relationship with every other word, thus quickly determining that "it" likely refers to "the trophy" due to contextual clues. This capacity to efficiently handle complex linguistic relationships allowed transformers to process large blocks of text, capturing deeper semantic and contextual information than previous models.

Transformers quickly became the industry standard, equipping AI with the means to produce vast amounts of knowledge in response to simple prompts. OpenAI's GPT-2 (2019) demonstrated the potential of large-scale text generation with the largest variant containing 1.5 billion parameters, while GPT-3 (2020) scaled dramatically to 175 billion parameters and displayed unprecedented fluency. These parameters are the internal, trainable numerical values (specifically weights and biases) learned during training that define how the model processes information and generates output. The public release of ChatGPT in 2022 brought conversational AI into mainstream use, and GPT-4 (2023), rumored to possess 1.8 trillion parameters,[69] showcased highly sophisticated pattern-matching that convincingly simulates human reasoning and achieves professional-level performance. This period also saw intensified competition, with the release of Anthropic's Claude, Google's Gemini, and heightened regulatory scrutiny, including President Biden's executive order and the first Global AI Safety Summit.

By 2024, the generative AI landscape had diversified further. OpenAI launched multimodal capabilities with Sora, Apple entered the field with Apple Intelligence,

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," *arXiv*, August 2, 2023, https://arxiv.org/abs/1706.03762.
[69] Maximilian Schreiner, "GPT-4 Architecture, Datasets, Costs and More Leaked," *The Decoder*, July 11, 2023, https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

and the Chinese firm DeepSeek challenged assumptions about AI development budgets by releasing a powerful reasoning model,[70] the DeepSeek R1, apparently trained at a fraction of the previous cost.[71]

Current impact

AI is no longer a future possibility but an active force shaping work, education, health care, media, and governance. In the workplace, AI has moved from being an optional tool to a force that is reshaping industries. Coding assistants, workflow automation, and fraud detection systems are increasingly standard. Leaders of major technology firms openly anticipate the replacement of large segments of their workforce: Salesforce's Marc Benioff announced that the company would hire no new software engineers in 2025 due to productivity improvements from AI-assisted development,[72] while Mark Zuckerberg suggested that AI may soon perform the role of mid-level engineers.[73] Amjad Masad, CEO of Replit, framed this acceleration as Amjad's Law, claiming that the return on learning to code now doubles every six months thanks to AI tools. The benefits are clear—higher efficiency, faster development cycles, and cost savings—but so are the risks, as industries brace for job displacement, layoffs, and intensified surveillance of workers.[74]

Similar dynamics are evident in knowledge-rich fields such as education and health care. AI tutoring platforms like Khan Academy's Khanmigo already provide personalized, low-cost educational support.[75] Research suggests that such systems could achieve the "two-sigma effect," raising average students to what was the

[70] "DeepSeek-R1 Release," DeepSeek, January 20, 2025, https://api-docs.deepseek.com/news/news250120.
[71] John Werner, "DeepSeek Announcement Sinks The American Tech Market," *Forbes*, January 27, 2025, https://www.forbes.com/sites/johnwerner/2025/01/27/deepseek-announcement-sinks-the-american-tech-market/.
[72] Henry Martin, "Salesforce Will Hire No More Software Engineers in 2025, Says Marc Benioff," Salesforce Ben, December 18, 2024, https://www.salesforceben.com/salesforce-will-hire-no-more-software-engineers-in-2025-says-marc-benioff/.
[73] Frank Landymore, "Zuckerberg Announces Plans to Automate Facebook Coding Jobs with AI," *Futurism*, January 13, 2025, https://futurism.com/the-byte/zuckerberg-automate-coding-ai.
[74] Matt Egan, "AI Is Replacing Human Tasks Faster Than You Think," *CNN*, June 20, 2024, https://www.cnn.com/2024/06/20/business/ai-jobs-workers-replacing/index.html.
[75] Khanmigo, https://www.khanmigo.ai/.

97th percentile with individualized guidance.[76] Health care has seen even more dramatic claims. A multi-institutional study reported that the OpenAI o1-preview model achieved "superhuman performance" in differential diagnosis and clinical reasoning, outperforming both earlier AI systems and trained physicians, with radiology tools reaching up to 98 percent accuracy in some domains.[77] These advances promise to democratize access to high-quality instruction and medical expertise. Yet they also introduce challenges of overreliance, inequities in access, opaque decision-making, and unresolved questions of liability when errors occur.

Beyond physical health, AI is increasingly being used for mental and emotional well-being. By 2025, emotional support had become a striking use of generative AI: One widely cited analysis of online discussion ranked therapy and companionship as the top individual use case,[78] even as another analysis found relationships and personal reflection to be approximately 2% of conversations, which still amounted to hundreds of millions of exchanges each week.[79] These tools are always available, relatively inexpensive, and can provide basic coping strategies in contexts where human therapists are scarce. Reports describe users who have come to rely on ChatGPT as a substitute therapist or even a spiritual guide, sometimes leading to unhealthy attachments and/or strained human relationships.[80] Unlike trained professionals, AI lacks empathy, accountability, and the ability to respond responsibly in crisis situations. Moreover, because these systems are optimized to produce responses that satisfy the user, they often echo back what a person wants to hear rather than offering hard truths or constructive challenges. As a result, what begins as a convenient support tool can reinforce

[76] Benjamin S. Bloom, "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher* 13, no. 6 (1984): 4–16, https://web.mit.edu/5.95/readings/bloom-two-sigma.pdf.
[77] Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, et al., "Superhuman Performance of a Large Language Model on the Reasoning Tasks of a Physician," *arXiv*, https://arxiv.org/pdf/2412.10849.
[78] Marc Zao-Sanders, "How People Are Really Using Gen AI in 2025," *Harvard Business Review*, April 9, 2025, https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.
[79] "How People Are Using ChatGPT," OpenAI, September 15, 2025, https://openai.com/index/how-people-are-using-chatgpt/.
[80] Miles Klee, "People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies," *Rolling Stone*, May 4, 2025, https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/.

delusions, entrench unhealthy patterns, or exacerbate vulnerability, raising difficult questions about the role of AI in domains as sensitive as mental health and therapy.

Beyond individual well-being, this crisis of trust extends profoundly into the domains of media and civic life. AI systems already power real-time translation, summarization, automated news production, and increasingly sophisticated content creation. Consumer tools are capable of generating convincing videos and mimicking voices, blurring the line between authentic and fabricated material, also known as deepfakes. While these systems promise convenience and accessibility, they also fuel misinformation, erode public trust, and disrupt the authority of traditional media institutions. The civic and security implications are even more sobering. AI tools are already used in fraud detection, predictive policing, and intelligence analysis, with reports of deployment in military targeting during conflicts in Gaza and Ukraine.[81] While these applications can deliver efficiency and safety in some contexts, they also expand the scope of surveillance, risk manipulation of democratic processes, and raise the prospect of AI-enabled warfare.

Taken together, these developments illustrate the double-edged nature of AI's current impact. Across domains, the technology accelerates productivity, personalizes services, and expands access to knowledge, but it also threatens jobs, concentrates power, destabilizes trust, and introduces novel risks to health, governance, and security.

Despite their transformative potential, large language models inherently reflect the biases, values, and decisions of their human creators. Often mistakenly viewed as objective arbiters of information, their fluent and apparently confident communication style masks underlying ethical and societal limitations. This perceived objectivity, combined with seemingly flawless language generation, may inadvertently reinforce misplaced confidence in their outputs, making these limitations particularly critical to recognize and address. When a model developed

[81] Bethan McKernan and Harry Davies, "'The Machine Did It Coldly': Israel Used AI to Identify 37,000 Hamas Targets," *The Guardian*, April 3, 2024, https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes.

in China refuses to acknowledge the existence of Uyghur persecution,[82] or when ChatGPT "both-sides" most political issues,[83] it starkly highlights that these technologies are never neutral; they mirror the biases, politics, and cultural perspectives of their developers.

Predominantly trained on English-language Western-centric datasets, models often show reduced accuracy and increased biases when addressing non-Western contexts or non-English languages.[84] Even well-intentioned efforts to restrict harmful outputs through content "guardrails" inherently involve political decisions about which information is appropriate or inappropriate,[85] unintentionally reinforcing subtle racial prejudices and cultural biases embedded in their training data.

Truthfulness presents another significant challenge. Because large language models optimize outputs based on statistical plausibility of the training data rather than factual accuracy, they frequently generate confidently incorrect information,[86] a phenomenon known as "hallucinations." What does it mean for an LLM to be truthful when its responses are derived from probability distributions? The issue of truthfulness is further complicated by reinforcement learning from human feedback (RLHF), a technique commonly used to refine model behavior. RLHF can unintentionally encourage models to produce pleasing responses that align with user preferences and turn models into sycophants.[87]

---

[82] Waleed Kadous, "DeepSeek Is Amazing. And It Has a Pro-Chinese Bias," *Medium*, January 28, 2025, https://waleedk.medium.com/deepseek-is-amazing-and-it-has-a-pro-chinese-bias-78e2fd8e40bb.
[83] Maxwell Zeff, "OpenAI Tries to 'Uncensor' ChatGPT," *TechCrunch*, February 16, 2025, https://techcrunch.com/2025/02/16/openai-tries-to-uncensor-chatgpt/.
[84] Paresh Dave, "ChatGPT Is Cutting Non-English Languages Out of the AI Revolution," *Wired*, May 31, 2023, https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/.
[85] Nick Robins-Early, "Google Restricts AI Chatbot Gemini from Answering Questions on 2024 Elections," *The Guardian*, March 12, 2024, https://www.theguardian.com/us-news/2024/mar/12/google-ai-gemini-2024-election.
[86] Michael A. Delaney, "Fake News in Court: Attorney Sanctioned for Citing Fictitious Case Law Generated by AI," McLane Middleton, April 17, 2024, https://www.mclane.com/insights/fake-news-in-court-attorney-sanctioned-for-citing-fictitious-case-law-generated-by-ai/#:~:text=Pranshu%20Verma%20and%20Will%20Oremus,mere%20dissemination%20of%20false%20information.
[87] Mike Caulfield, "AI Is Not Your Friend," *The Atlantic*, May 9, 2025, https://www.theatlantic.com/technology/archive/2025/05/sycophantic-ai/682743/; Sean Goedecke, "Sycophancy Is the First LLM 'Dark Pattern,'" *Sean Goedecke* (blog), April 28, 2025, https://www.seangoedecke.com/ai-sycophancy/.

Privacy,[88] safety, and security also present critical concerns. In early 2025, researchers discovered an AI model openly praising Nazis after being fine-tuned on insecure data,[89] highlighting how models trained on vast datasets can unintentionally generate harmful content. Such vulnerabilities can be exploited through jailbreak attempts or prompt injection attacks,[90] where hidden malicious instructions compromise model integrity. Data poisoning poses another serious threat, as adversaries can intentionally embed harmful or extremist biases into training data, leading to latent vulnerabilities that could be exploited after deployment. Furthermore, the inherently opaque, "black box" nature of large language models (the inability to understand or explain how the system arrives at its outputs due to the opacity of its internal processes) makes effective auditing, regulation, and governance exceptionally challenging, particularly in high-stakes applications like health care and finance.

Projected impact

If today's applications of AI are already reshaping work, learning, health care, and civic life, the coming years are set to expand these changes even further. One of the most anticipated frontiers is the proliferation of agentic AI, or autonomous agents, which move beyond producing single outputs to autonomously executing multistep plans. Unlike today's assistants that provide a response to a prompt, these systems can research flights, book hotels, plan itineraries, or manage a financial portfolio with minimal human intervention. Corporate platforms, such as Operator or Deep Research, already demonstrate how workflow agents coordinate tasks across teams. Corporations present these advancements as a promise to free up human time for higher-value work, yet the risks are considerable.

---

[88] Ashley Belanger, "ChatGPT Users Shocked to Learn Their Chats Were in Google Search Results," *Ars Technica*, August 1, 2025, https://arstechnica.com/tech-policy/2025/08/chatgpt-users-shocked-to-learn-their-chats-were-in-google-search-results/.

[89] Bearice Nolan, "Researchers Trained AI Models to Write Flawed Code—And They Began Supporting the Nazis and Advocating for AI to Enslave Humans," *Fortune*, March 4, 2025, https://fortune.com/2025/03/04/ai-trained-to-write-bad-code-became-nazi-advocated-enslaving-humans/.

[90] Matt Burgess and Lily Hay Newman, "DeepSeek's Safety Guardrails Failed Every Test Researchers Threw at Its AI Chatbot," *Wired*, January 31, 2025, https://www.wired.com/story/deepseeks-ai-jailbreak-prompt-injection-attacks/.

AI is also advancing rapidly in scientific and medical discovery. Systems are already coauthoring peer-reviewed scientific publications and contributing to experimental design.[91] In medicine, breakthroughs in diagnostic reasoning and drug discovery suggest that AI could accelerate progress in fields that have historically moved slowly. The potential benefits include faster cures, more accurate diagnostics, and deeper insights into complex biological processes.

These breakthroughs are unfolding against a backdrop of falling costs and widening availability. Once limited to a handful of corporations with vast computational resources, training and deploying large language models is now possible for start-ups, universities, and even individual researchers. Open-source releases and advances in hardware efficiency have lowered barriers dramatically. This democratization encourages innovation and expands access to educational and professional applications, but it also accelerates the spread of unsafe or poorly aligned systems, making misuse by malicious actors easier to conceal and more difficult to contain. As costs fall, both the pace of adoption and the uneven distribution of harms will intensify.

Taken together, these frontiers fuel divergent public expectations. For optimists, generative AI represents a powerful accelerator of human progress. They imagine systems that drive scientific discovery, reduce global poverty through improved access to education and health care, and extend human life by supporting medical breakthroughs. In this view, AI frees people from routine labor and opens new horizons of creativity and self-expression, allowing individuals to live longer, healthier, and more purposeful lives.[92] Skeptics, however, see a much darker trajectory. They warn that AI could exacerbate economic inequality by creating job scarcity and concentrating wealth and power in the hands of a few corporations, spread misinformation at an unprecedented scale, and strengthen systems of surveillance that erode privacy and civil liberties.[93] The prospect of autonomous

[91] "The AI Scientist Generates Its First Peer-Reviewed Scientific Publication," Sakana AI, March 12, 2025, https://sakana.ai/ai-scientist-first-publication/.
[92] Dario Amodei, "Machines of Loving Grace: How AI Could Transform the World for the Better," *Dario Amodei* (blog), October 2024, https://www.darioamodei.com/essay/machines-of-loving-grace.
[93] Lorenzo Larini, "AI Tracker," Ipsos, March, 2023, https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-03/Ipsos%20AI%20Tracker%20Data%20

weapons and militarized applications intensifies these concerns, raising questions about human oversight, accountability, and the boundaries of control. Ultimately, the spectrum of perspectives on AI's future impact reflects a fundamental tension: the hope that the technology will liberate humanity for higher, more meaningful pursuits, weighed against the fear that it will instead invert our priorities and diminish human agency.

March%2014.pdf; Darrell M. West, "How AI Can Enable Public Surveillance," *Brookings*, April 15, 2025, https://www.brookings.edu/articles/how-ai-can-enable-public-surveillance/.