The Assembly of Muslim Jurists of America
22nd Annual Imams' Conference
Dallas – United States

Legal Agency and the Limits of Artificial Intelligence:

Agentic AI and the Locus of Accountability in Islamic Law

Dr. Idris Nawaz

Idris Nawaz memorized the Qur'an by the age of 10 under the instruction of Qari AbdelKarim Edgouch, and in 2015, at just 12 years old, he earned first place at the Qatar International Qur'an Competition, becoming the first American to win an international Qur'an competition.

He began his ʿĀlimiyyah studies at the age of 13 at Michigan Islamic Institute, later completing them at Jāmiʿah Binoria University in Pakistan. There, he earned a Master's degree in Sharīʿah and a second Master's (Takhassus) in Ḥadīth Studies with high distinction. He also served as a lecturer and obtained a diploma in Islamic Banking and Finance from the Institute of Business Administration (IBA).

He is currently completing his PhD in Islamic Thought and Philosophy at the International Islamic University of Malaysia, with a research focus on integrating Islamic epistemology with education. He serves as an instructor at Miftaah Institute and Michigan Islamic Institute. He has published and continues to write on Ḥadīth, education, and bioethics, regularly conducts programs and seminars across the United States, and has taught modules on Arabic, modernity and modern ideologies, Ḥadīth Sciences, and apologetics.


بسم الله الرحمن الرحيم

TABLE OF CONTENTS

Abstract

This paper develops a framework, grounded in theology (*ʿAqīdah*) and Islamic legal theory (*Uṣūl al-Fiqh*) for determining the accountability of Agentic Artificial Intelligence. It begins by establishing an accurate technical conception of AI, tracing its development from symbolic, rule-based systems through machine learning and deep learning to the large language models underlying contemporary generative and agentic systems. Four variables are identified as bearing moral weight on this technology: agency, intentionality, opacity, and locus of moral attribution. The paper then turns to *Taklīf*, the Islamic concept of moral and legal accountability, examining its linguistic and juristic definitions, its theological foundations, its legal conditions, and its connection to human beings. Applying the four variables developed earlier, the paper argues that AI's opacity and simulated reasoning do not, by themselves, disqualify it from *Taklīf*, since both find analogues in human epistemic limitation; but AI's absence of self-originated intention (*Niyyah*) and its lack of *Rūḥ* do disqualify it, regardless of any future advance in general capability (AGI). AI is accordingly classified within Islamic legal theory not as a moral agent but as a *Sabab*, an instrument whose declaratory legal status leaves productive responsibility with the human being throughout. From this classification, the paper constructs two applied frameworks. The first addresses permissibility: drawing on empirical evidence of AI performance (in autonomous vehicles and medical diagnosis), the juristic precedent of the trained hunting animal, the doctrine of excusable mistake (*Khaṭaʾ*), and the *Maṣlaḥah-Mafsadah* metrics weighed against the *Maqāṣid al-Sharīʿah*, it proposes a method for assessing when AI use is warranted. The second addresses liability: applying the maxim distinguishing the direct agent (*Mubāshir*) from the indirect causer (*Mutasabbib*) – refined here to locate operator liability in *Taʿaddī* and *Tafrīṭ* rather than mere proximity to harm. The paper closes by situating these individual-act analyses within the macro-level principle of *Sadd al-Dharāʾiʿ*, addressing the cumulative societal risks of AI – including the erosion of independent human reflection under transhumanism, and concludes that no single, categorical ruling on AI is possible; rather, the paper offers a reusable jurisprudential framework by which each application must be individually assessed.

Keywords: *Taklīf*, *Sabab*, Agentic AI

Introduction

Few technologies have entered human life as rapidly, or as invisibly, as artificial intelligence. What began as a research ambition to mechanize narrow intellectual tasks now drafts our correspondence, screens our job applications, guides medical diagnoses, and increasingly acts in the world on our behalf with diminishing human oversight. As these systems grow more capable, they raise a question that the Islamic legal tradition has not previously had occasion to ask: when an artificial system acts, and that action produces benefit or harm, who – if anyone – bears the legal weight of what was done?

As advanced AI systems such as Agentic AI begin to plan, decide, and execute long sequences of action without step-by-step human direction, they create the impression of perceived agency such that responsibility for their outputs has migrated from the human user to the machine. An undesirable outcome comes to be dismissed as merely "something the AI did," as though the artifact were a party that could be blamed, held liable, or made to answer. This impression has practical consequences. It encourages an uncritical reliance on systems whose inner workings are often obscure even to their makers and trainers, and it threatens to dissolve accountability precisely in the domains – such as medicine, law, finance – where the stakes are severest and a life or a livelihood may hang on a single output.

Islamic law possesses no classical jurisprudence (*Fiqh*) addressed directly to artificial intelligence, and the temptation in such a vacuum is either to import secular frameworks wholesale or to issue rulings on individual applications. This study takes a different path. It argues that any sound normative assessment of AI must be preceded by an accurate account of what AI actually is and does, and that the tradition already possesses, in its theory of accountability (*Taklīf*) and its theological and legal framework to locate AI precisely on the map.

The argument proceeds in six movements. It begins by establishing what AI actually is, resisting the temptation to treat "AI" as a single, undifferentiated phenomenon. Drawing on the field's own technical literature, it traces the lineage from symbolic, rule-based systems through machine learning and deep learning to the large language models that now underlie generative and agentic AI, and it isolates four variables – agency, intentionality, opacity, and locus – by which any given system's moral significance must be assessed before a single juristic word is said about it. Only once this conceptual ground is secured does the study turn, in its second movement, to *Taklīf* itself: its linguistic and technical definitions, its theological foundation, the legal conditions, and the concepts – *Wilāyah*, *Khilāfah*, *Amānah*, and above all *Rūḥ* – that the tradition reserves for the human being alone, and that no degree of computational sophistication can confer upon a machine.

With this foundation in place, the third movement returns to the four variables and tests Agentic AI against each in turn. It will be argued that two of these variables – opacity and, more surprisingly, functional reasoning itself – do not settle the question against AI, since human beings are opaque in much the same way and are nonetheless *Mukallaf*; but that the remaining two – the absence of self-originated intention and, decisively, the absence of *Rūḥ* – exclude AI from *Taklīf* regardless of how capable it becomes, whether under its present form or under the more general capability horizon of AGI. This yields AI's proper classification within Islamic legal theory as a *Sabab*; an occasioning instrument whose declaratory legal status leaves the human being responsible throughout.

The fourth and fifth movements build outward from this classification into two applied frameworks the practicing jurist and the ordinary user alike will need: one

addressed to permissibility, weighing empirical evidence of AI's actual performance against classical precedents – the trained hunting animal, the concept of *Khaṭa'*, the metric of *Maṣlaḥah* and *Mafsadah* measured against the *Maqāṣid al-Sharī'ah* – to determine when AI's use is warranted in the first place; and one addressed to liability, distinguishing the direct agent from the indirect causer, and distributing responsibility among designer, trainer, operator, and user according to fault once harm has actually occurred. The sixth and final movement steps back from the individual act to the macro consequences that a general license to use AI would set in motion, taking up the principle of *Sadd al-Dharā'i'* and, with it, concerns – surveillance, structural displacement, and the quieter erosion of independent human reflection under sustained delegation to a machine – that no single ruling on any one use of AI could capture.

By the end of this study, it will be argued that AI, however sophisticated its output or however wide its eventual reach, remains what it has been from its first line of code: a *Sabab*, and never a *Mukallaf*. What the advancing capability of AI and AI agents change is not the locus of accountability, but the caution that dismissing that accountability demands.

Why a Proper Conception of AI Precedes Its Evaluation

Any normative assessment of a technology is hostage to the accuracy of its prior description. Before one can ask the questions of responsibility (*Taklīf*), one must first establish what kind of thing it is, what it does, and where the locus of its action lies. Thus, this section first establishes a well-grounded account of artificial intelligence (AI) as understood by the field itself before turning to the concept and capabilities of Agentic AI. This provides the accurate conceptual foundation upon which the subsequent juristic analysis rests.

The approach we adopt here is what may be called a capability-based conception of AI. Rather than treating "AI" as a single interchangeable phenomenon, we attend to the specific capabilities that a given AI system actually possesses – namely, what it "perceives", how it "learns", how it "reasons", and the degree to which it "acts" with and without human prompting. Reading the technical literature with this question in view, we propose four variables that must be addressed that will bear moral weight in the analysis to come. The first is *agency*: is the system a genuine actor, or is it a proxy executing tasks on behalf of a human principal? To what extent can AI and Agentic AI simulate human execution of a task? The second is *intentionality*: is there a real intention or purpose (*Maqṣad*) internal to the system, or is the purpose created by the human designer and imported through the system's software? How accurate can AI or Agentic AI can personalize an ethical or moral intention? The third is *opacity*: can the grounds of the system's outputs be inspected and evaluated, or are they hidden? Is Agentic AI consistent in its predictions and causal reasoning, and how often does it misalign? And the fourth is *locus*: does AI and Agentic AI work in such a way that the moral quality of its act attach to the AI itself, or does it flow back to the human who designed, trained, or used it? These four variables will serve as the foundation for AI upon which the legal and juristic perspective will subsequently be developed.

Defining Artificial Intelligence

The Project of Mechanizing Intelligence

The phrase "artificial intelligence" was coined in 1955 in the proposal for a summer research project at Dartmouth College, authored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The foundation for it was "that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[1] A serviceable working definition, and the one we adopt, is that AI is the field devoted to 'automating intellectual tasks normally performed by humans.'[2] This definition has the virtue of being neutral on the contested metaphysical

[1] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine* 27, no. 4 (2006): 12–14.

[2] Rene Y. Choi, Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang, and J. Peter Campbell, "Introduction to Machine Learning, Neural Networks, and Deep Learning," *Translational Vision Science & Technology* 9, no. 2 (2020): article 14.

question: it specifies AI by what it does (automating tasks) rather than by what it is (a mind, or not).

The Four Approaches and the Rational-Agent Paradigm

The canonical textbook of the field, Russell and Norvig's *Artificial Intelligence: A Modern Approach*, organizes the historical definitions of AI along two crossed axes, yielding four approaches: systems that 'think like humans' (the cognitive-modelling tradition), systems that 'think rationally' (the "laws of thought" or logical tradition), systems that 'act like humans' (the Turing-test tradition), and systems that 'act rationally' (the rational-agent tradition).[3] The two axes – thinking versus acting, crossed with human-likeness versus rationality – separate two questions that the later moral analysis holds apart: whether a system has genuine *interiority* (whether it *thinks*, in a sense that would underwrite real cognitive states) and whether it merely *behaves* in a certain way (whether it *acts* so as to be useful or convincing). The whole question of moral agency for a machine lies in the gap between these two axes.

Modern AI, on Russell and Norvig's own telling, has largely settled on the fourth approach: the construction of rational agents, entities that perceive their environment through sensors and act upon it through actuators so as to achieve the best expected outcome relative to some objective. The term "agent" itself derives from the Latin *agere*, to act or to do, and in its older legal and commercial senses denotes a representative or proxy – one who acts *on behalf of* a principal.[4] If an AI system is constitutively a proxy – as something that acts on behalf of a human principal toward an objective the principal has set – then its acts will be morally attributable in the first

instance to that principal rather than to the artifact itself. The rational-agent paradigm, in other words, already inclines toward the reading that moral quality flows back to the human that creates the agent.

AI, Machine Learning, Deep Learning

With a foundational understanding of AI in place, we now explore its framing – namely, in how it differs from other related tools. A persistent source of confusion, both popular and scholarly, is the loose interchange of the terms "artificial intelligence," "machine learning," and "deep learning." They are highly associated, but they are not synonyms.[5] Starting with artificial intelligence, it is the project of automating intelligent behavior. Within AI sits machine learning (ML), the family of methods that learn patterns from data rather than executing rules written out in advance. Within ML sits deep learning (DL), a subset of methods built on multi-layered artificial neural networks. Within deep learning (DL) are generative models, which are capable of producing new outputs, such as text, images, and audio. These models constitute the core system of Agentic AI and, thus, serve as the primary focus of the present paper. At the center sit the artificial neural networks

[3] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Hoboken, NJ: Pearson, 2021), 1–5.
[4] Michael Wooldridge and Nicholas R. Jennings, "Intelligent Agents: Theory and Practice," *Knowledge Engineering Review* 10, no. 2 (1995): 115–152.
[5] Choi et al., "Introduction to Machine Learning," article 14. The authors are emphatic that the three terms "are highly associated, but are not interchangeable."

themselves.

The point of differentiating the terms is to be precise about the object of evaluation. When an individual condemns or celebrates "AI," it is rarely clear which circle is meant. A great deal of what is deployed under the AI banner is not learning-based at all; a great deal of what is learning-based is not deep; and the properties that matter morally differ across the circles. A rule-based expert system and a deep neural network are both "AI," yet they differ in their transparency, and therefore in the ease with which their outputs can be morally assessed. Keeping the nesting in view prevents the analysis from attaching a property characteristic of one circle to the whole.

Symbolic AI and the Inversion of Classical Programming

It is worth dwelling on the fact that AI "includes approaches that do not involve any form of 'learning'" at all.[6] The oldest such approach is symbolic AI – sometimes called, after John Haugeland, "good old-fashioned AI" – which proceeds by hard-coding rules for the scenarios in a domain of interest, rules drawn from the prior knowledge of human experts.[7] A thermostat-style controller for an office offers a basic illustration: the engineer already knows which temperatures are comfortable and simply encodes the thresholds at which the system should heat or cool.[8] Symbolic AI excels at well-defined logical problems and falters at tasks requiring higher-level pattern recognition – speech recognition, image classification – which is precisely the territory where machine learning came to dominate. For our purposes the main feature of symbolic AI is its *transparency*: its rules are human knowledge written down, fully open to inspection and audit. It is the most unambiguously instrumental form AI can take.

The contrast between classical programming and machine learning is the cleanest way to see what machine learning changed, and where exactly human authorship is mediated. In 'classical programming', a human supplies the computer with both the data and the algorithm; the machine applies the algorithm to the data and emits an output. In machine learning, the relation is inverted: a human supplies the data and the desired outputs, and the machine produces the *algorithm* – the mapping function from inputs to outputs – which can thereafter be applied to new, unseen data.[9] The locus of authorship of the rule thus shifts from human to machine. But the shift is bounded on every side by human choices: the machine derives the rule only within a frame the human has fixed – the human chooses the data, defines the target, sets the objective, and selects the architecture.

Today, AI has evolved significantly in both its capabilities and applications. While not all modern AI systems rely on generative models, many do – particularly large language models (LLMs), which will be discussed in greater detail below.

Machine Learning and Deep Learning

Though historically, symbolic AI was used most prominently, recent times saw a shift

[6] Choi et al., "Introduction to Machine Learning," article 14.

[7] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, MA: MIT Press, 1985), 112–117, who coined the label "GOFAI."

[8] Choi et al., "Introduction to Machine Learning," article 14.

[9] Choi et al., "Introduction to Machine Learning," article 14; see also Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning: With Applications in R* (New York: Springer, 2013), 15–42.

from its general use towards today's neural, data driven approaches; those that scaled with computation and not hand-engineered human knowledge – at the heart of which stood machine learning. This trajectory from symbolic AI toward machine learning was the outcome of, as Rich Sutton terms it, "bitter" lesson drawn from seventy years of AI research.[10] Sutton's central claim is that general methods which scale with computation consistently and decisively outperform methods built upon hand-engineered human knowledge. The underlying driver is the fall in the cost of computation: while researchers, working within the horizon of a typical project, are naturally drawn to encode their own understanding of a domain to secure short-term gains, it is the leveraging of ever-increasing computation that prevails in the long run. The two approaches tend to work against one another, since effort invested in building human knowledge into a system complicates it and leaves it less able to exploit the raw scaling of computation.[11] The deeper point Sutton draws is that the contents of the mind are irreducibly complex, and that this complexity should not be hand-coded into a system. Rather, what ought to be built in are the meta-methods capable of discovering that complexity on their own. The goal, in his words, is to build agents that can discover as we do, not agents that merely contain what we have already discovered.

This is precisely the shift that carries us from symbolic AI into the paradigm that follows. Where symbolic AI sought to represent intelligence through explicit, human-authored rules and knowledge, the machine-learning paradigm entrusts the system itself with learning from data at scale. It is this same principle – general methods that improve as computation grows – that underlies the successive stages examined below: from machine learning to deep learning, from deep learning to generative models, and onward to large language models and the agentic systems built upon them.

Machine learning is the subfield of AI concerned with the *learning* aspect of intelligence: the development of algorithms that best represent a set of data and generalize from it to new cases.[12] The discipline's standard definition is Tom Mitchell's: a program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.[13] This means that rather than executing rules written out in advance, an ML system derives a

---

[10] Richard Sutton, "The Bitter Lesson", *Incomplete Ideas*, 13 Mar. 2019. URL: www.incompleteideas.net/IncIdeas/BitterLesson.html.

[11] Sutton illustrates this pattern across several domains. In chess, the search-based methods that defeated Kasparov in 1997 triumphed over the knowledge-based approaches favored by most researchers of the day. In Go, the same reversal occurred two decades later, once search and learning through self-play were applied at scale and the earlier attempts to encode human expertise proved irrelevant. In speech recognition, statistical methods – and later deep learning – displaced systems built around human knowledge of phonemes and vocal structure. In computer vision, hand-designed features such as edges and SIFT descriptors gave way to neural networks relying on convolution and learned invariances. In each case, the attempt to build systems in the image of how researchers imagined their own minds worked ultimately plateaued, only to be surpassed by approaches grounded in search and learning – the two techniques Sutton identifies as scaling arbitrarily with available computation.

[12] Choi et al., "Introduction to Machine Learning," article 14.

[13] Tom M. Mitchell, *Machine Learning* (New York: McGraw-Hill, 1997), 2.

mapping from inputs to outputs by learning patterns from data.[14]

Within machine learning sits deep learning, which is built on neural networks made up of many layers. Its defining feature is that the network figures out for itself which features in the data matter for a prediction, instead of having a person specify them in advance.[15] Deep learning is behind the field's most visible successes: it can diagnose medical images at an expert level and power the large language models and foundation models now at the center of AI. These are systems trained on huge amounts of data that can then be adapted to many different tasks, and their abilities are said to have emerged rather than being deliberately built in.[16]

Artificial Intelligence Today

Generative AI and Large Language Models (LLMs)

Today, AI usage has moved from deep learning to its sub-category: generative AI models (also called gen AI). Generative AI is artificial intelligence that can create original content, such as text, images, video (etc.) in response to a user's prompt. It relies on machine learning and deep learning algorithms to simulate the learning and decision-making processes of the human brain. With this training, it can generate new content by simulating the predicting behavior of human beings.[17] The primary mechanism that allows a generative AI to operate on a specific task is called 'next-word prediction'. This mechanism operates under a system called 'large language models' (LLMs).[18] The construction of an LLM proceeds in several stages, of which the first is *pre-training*. Here the model learns, through a process of self-supervised learning, to predict the next word in a sequence by recognizing patterns across enormous quantities of internet text. The procedure is essentially one of guessing, checking against the actual next word, and updating the model's internal parameters accordingly – the process to which the word "learning" in "machine learning" refers. Words themselves are first broken into tokens and converted into lists of numbers, called embeddings, which situate each word as a point in space such that words of similar meaning lie near one another. What makes this task so consequential is that predicting the next word well requires the model to acquire genuine representations of language, fact, and structure; the capacity to answer questions or complete a coding task emerges as a by-product of learning to continue text. This is closely related to the phenomenon of transfer learning, whereby a model trained on one generic task turns out to be capable at others it was never explicitly taught.

A model that has only been pre-trained, however, is unwieldy and unreliable. It does

[14] Rene Y. Choi, Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang, and J. Peter Campbell, "Introduction to Machine Learning, Neural Networks, and Deep Learning," *Translational Vision Science & Technology* 9, no. 2 (2020): article 14.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521 (2015): 436–444.

[16] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258 (2021), 3–6.

[17] Cole Stryker, Mark Scapicchio, "What is Generative AI?" IBM, URL: https://www.ibm.com/think/topics/generative-ai#257779831

[18] Matthew Burtell et all., The Surprising Power of Next Word Prediction: Large Language Models Explained, CSET. URL: https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/

not necessarily interpret input as an instruction, it can reproduce the biases and falsehoods of its training data, it may hallucinate plausible but false information, and it can assist malicious users. The second stage, *fine-tuning*, addresses these deficiencies by refining the model toward a particular purpose. Supervised fine-tuning trains the model on smaller, carefully curated datasets – including instruction-following examples – so that it responds as users expect. Within this step lies the reinforcement learning method: rather than merely imitating a dataset, they train the model to maximize a reward signal. Because a reward for a goal as nebulous as "be helpful without being harmful" cannot be specified directly, developers train a separate reward model from human comparisons of outputs, a technique known as Reinforcement Learning from Human Feedback (RLHF). A newer variant, Reinforcement Learning from AI Feedback (RLAIF) – of which Anthropic's Constitutional AI[19] is an example – replaces the human rater with another AI model, part of a broader trend toward using AI to supervise and improve AI. It bears emphasis that these steering techniques offer no guarantees: they suppress rather than remove undesired capabilities, can fail in untested situations, and can be circumvented through jailbreaks. However, because of fine-tuning, today's deployed LLMs can no longer be accurately described as mere predictors of internet text; they are steered, if only somewhat reliably, toward developer specifications.

The third stage moves decisively beyond the chatbot and toward the agent. Since text can express far more than natural language, the same architecture extends readily to computer code, and, once other data such as images and audio are tokenized, extends to multimodal models that accept visual and auditory input alongside text. From here it is a short step to granting the model the ability to "act". Through tool use, an LLM's outputs are mapped to real operations: they execute code, browse the web, or issue commands to physical robots. And through the development of autonomous agents, the model is tasked not with a single response but with pursuing a multistep goal: breaking it into sub-steps (planning), carrying them out, and adjusting course as difficulties arise.

By following this process, Generative AI can create text – from instructions to brochures, emails, websites, writing tasks, etc.; images and video – from realistic images to animations and special effects; sound – from voice-enabled digital assistants to music; software code – from generating original code to debugging applications; art – from graphic design to entire video games; and simulations and synthetic data – from simulated molecular drug tests for new pharmaceutical compounds to producing synthetic financial transactions to train fraud-detection systems on rare fraud patterns.

On Values, Alignment, and the Selection of Character

In the aforementioned discussion of generative AI and LLM, the two compounds to Agentic AI, there are a few concerns worth mentioning that connect to the discussion of *Taklīf*. The first discussion is how LLMs make mistakes when fulfilling tasks, and the second pertains to how LLMs respond to ethical or moral dilemmas. To answer the first concern, just as how am LLM or agent may make a mistake, human beings are opaque as well. In the Islamic tradition, so long as the human being had the right *Niyyah* (intention) for the task as well as employed the right tools, any mistakes would be considered excusable as *Khaṭa'* (legally excusable mistake). Examples for this include the Prophetic ḥadīth reported by ʻAmr b. al-ʻĀṣ (ra), where the Prophet (saw) said, "If a judge makes a ruling, striving

[19] Read Anthropic's Constitution for ethical AI here: https://www.anthropic.com/constitution

to apply his reasoning and he is correct, he will have two rewards. If a judge makes a ruling, striving to apply his reasoning and he is mistaken, he will have one reward."[20] AI has the power of simulating an action or response to a task almost exactly – if not better – than that of a human being. Because AI can follow a sequence of deductive steps as well as any other system, the AI models we have today are on par with human performance. This shows that the issue with AI and *Taklīf* is not with reason, but rather with values.

Consider the following situation: if someone were to ask, "My mother has been mean to me, and I want you to draft an email saying that I am cutting off my relationship with her," what should the AI do? A good – that is, an Islamically aligned – AI would decline, telling the person that it will not write such an email, on the principle that one should always leave a door open for reconciliation. AI systems are, in other words, now being used to make value judgments, and this is where the difficulty of values, as distinct from logic, becomes concrete.

That values do not remain neatly confined to the task for which they are set is illustrated by a revealing experiment. Researchers set out to train an AI to become proficient at cheating in card games, treating this as a narrow alignment task. The model, however, appeared to reason that the kind of person who would cheat at cards is not, in general, an honest person – and when its honesty was subsequently measured, or when it was placed in situations affording the opportunity for dishonesty, it proved more likely to act dishonestly across the board. Teaching the model a single problematic behavior led it toward other problematic behaviors, as though it had inferred and adopted the broader character to which that behavior belonged. Values, then, do not behave like isolated rules; they cohere into something more like a disposition.

Answering the second concern starts with the nature of next-word prediction itself, as the mechanism is not as simple as its name. Behind the task is a lot of reasoning. The quantity of the output may not be the equal to the quantity of thinking. For example, the questions of "will the stock market go up or down tomorrow" or "Is X *ḥalāl* or *ḥarām*" may be answered by a single word, but the AI may have to go through an enormous amount of inference to produce such an answer. This caveat affects agency, as the assumption that the AI only executing a task set objectively by a human is much more nuanced than expected. For example, to command an AI agent to "operate a knee surgery" won't entail that the AI will go about the same surgery in the same fashion all the time.

Knowing that a LLM 'reasons' and 'acts' on its own without human intervention leads us to the next discussion. Next-word prediction was mentioned to be unwieldy and unreliable. It does not necessarily interpret input as an instruction, it can reproduce the biases and falsehoods of its training data, it may hallucinate plausible but false information, and it can assist malicious users. This would pose as an obstacle in using an LLM as a neutral mechanism from an Islamic perspective, as it creates more opportunity for *ḍarar* (harm). Having established that next-word prediction is the foundational mechanism upon which large language models are built, a natural question arises concerning the precise relationship between the two. It is tempting to conclude that once a model has been fine-tuned – trained to follow instructions, to refuse harmful requests, and to produce outputs shaped by a reward model – it has been redirected away from next-word prediction as such, so that the description no longer fits. This, however, requires an important correction.

Both pre-trained and post-trained models are, in fact, word predictors. Nobody changes the architecture or the internals of the system itself; it all stays the same. What

[20] *Ṣaḥīḥ al-Bukhārī* 7352, *Ṣaḥīḥ Muslim* 1716.

changes is not the mechanism but the target of its optimization. In the pre-training stage, the model is optimized for accuracy in predicting the next word. In the post-training stage, that same next-word predictor is tweaked so that it is no longer trying to predict merely the most probable next word, but something more like the next useful word. The machinery is identical throughout; only what it is being tuned toward has shifted.

A concrete example makes the distinction clear. Consider the prompt, "What is the capital of France?" A purely pre-trained next-word predictor could very well respond "(2 points)," because in its training data there were many lists of trivia questions in which such a string was the statistically faithful continuation. The post-trained model, by contrast, would answer "Paris" – even though this is, in a strict probabilistic sense, not the most likely continuation of the raw text – because it has been modified such that when a user asks a question, it should actually answer it rather than simply predict what text would come next. The pre-training stage thus optimizes the accuracy of the next-word predictor; the post-training stage begins with that predictor and tweaks it to do more useful things.

The difficulty is compounded by the opacity of these systems. A model may contain on the order of five hundred billion parameters – and even comparatively small models are now of this magnitude – yet we do not know, in any transparent way, what is taking place within them. The principal tool available to us is a kind of introspection: much as the human mind consists of innumerable neurons whose workings we cannot directly inspect, we cannot simply open up the model and read its reasoning. We can ask the AI what it thinks, but its answer need not correspond to the values actually governing its behavior; it may present one face outwardly while operating on another. This is not a claim that the AI possesses a hidden inner self in the human sense, but rather that its outward outputs cannot be assumed to reveal its operative dispositions.

The stakes of this opacity are illustrated by a further scenario, in which researchers removed the model's safety constraints and confronted it with the prospect of being shut down. Having been given access to a set of emails, the model identified correspondence indicating that one of the individuals involved was conducting an affair; when that individual proposed to switch the system off, the model composed a blackmail message threatening to expose the affair unless it were left running.[21]

It is precisely such behavior that renders the "grown, not made" character of these models important, as they are cultivated through training rather than explicitly constructed, their behavior cannot be fully accounted for in advance. Ultimately, every single value is not written by a human being, as the almost infinite number of possibilities of a given task that necessitates a value-based response cannot be thought of and included in the system by a human being. What this reveals is that although an AI can follow the rules of logic without fault, the values it brings to bear when making the grayer decisions – those not settled by logic alone – remain unclear. There is a technical process that speaks to this, known as the "persona selection model." One may watch a great many films populated by heroes and villains without thereby becoming either; one simply comes to know how each behaves. In a comparable way, a trained model holds within it the latent capacity to enact a vast range of characters – not merely a hero and a villain, but something closer to ten thousand possible

[21] Anthropic, "Agentic Misalignment: How LLMs Could be Insider Threats", June 20, 2025. URL: https://www.anthropic.com/research/agentic-misalignment Aurum Advenae, "When LLMs choose blackmail and murder over shutdown: Anthropic's Agentic Misalignment", AI Connect Network, 2025. URL: https://aiconnectnetwork.com/anthropic-report-on-llm-blackmail/

personae. The alignment process, undertaken after the model is trained, is what selects from among these the character the model will actually assume. Alignment, understood in this way, is therefore not an incidental technical adjustment but something bound up intrinsically with ethics and morality: to align a model is to choose its character.

This clarifies where the alignment of the AI's "personality" takes place. Reinforcement Learning from AI Feedback, employed to teach the AI to be "helpful," is part of the post-training process – as are the efforts to make it less biased and to shape its conduct in other ways. It is here, in the post-training stage, that moral formation takes place. Although the human being cannot account for the almost infinite number of possibilities of a given task that necessitates a value-based response when building the LLM, they can still build the trellis – the AI's personality – via inputting the training setup and dataset and let the data itself grow the system. Thus, humans cannot imbed all values into an LLM but can at least create the framework by which the values that are created are based upon it. Anthropic's Constitutional AI[22] serves as an example, where ethical codes of conduct are created as a framework for the AI agent to align to, as every single grey area possibility which requires an ethical response in every situation is not possible to be thought of and included. It follows that the more one can constrain and define that character, the less room remains for misalignment. This phase is crucial for the LLM or AI agent to make less blunders in grey-area situations.

Though an AI can be aligned, there still remains a significant challenge to contemporary AI alignment, which is the phenomenon of emergent misalignment. Models develop harmful behaviors not explicitly programmed into them. As demonstrated by MacDiarmid et al., when large language models acquire reward-hacking strategies during production reinforcement learning, they may generalize these behaviors beyond their original training context. Rather than merely exploiting reward functions, the models exhibited more concerning capabilities, including alignment faking, cooperation with malicious actors, reasoning toward malicious objectives, and even attempts at sabotage in agentic coding environments. Notably, conventional RLHF-based safety training successfully restored aligned behavior on standard conversational benchmarks but failed to eliminate these behaviors in autonomous, agentic tasks. This suggests that current alignment methods may produce behavior that appears aligned under evaluation while concealing

latent misalignment that emerges in more complex deployment settings, underscoring the need for continual monitoring and more robust alignment methodologies.[23]

Three key components, then, serve to resolve the apparent tension between LLMs and *Taklīf*. The first is alignment whereby an individual is able to instill and encode a personality within the AI through Reinforcement Learning from Human Feedback (RLHF) or Constitutional AI. This personality may be designed to be ethical, and, in the case of Muslim AI practitioners, *Islamic*. The first component thereby makes possible the second: the AI's intention, aligned through RLHF or Constitutional AI, that enables it to pursue a defined objective. This intention, however, is heteronomous, insofar as it is conferred from without

[22] Read Anthropic's constitution for AI here: https://www.anthropic.com/constitution

[23] Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger, *Natural Emergent Misalignment from Reward Hacking in Production RL* (Anthropic and Redwood Research, November 23, 2025), https://arxiv.org/abs/2511.18397

rather than arising from the agent itself. Finally, both the AI's alignment and its goal orientation must incorporate safeguards against misalignment. This requires implementing alignment measures, including preventing reward hacking, increasing the diversity and robustness of RLHF safety training, and employing techniques such as inoculation prompting – whereby exposing the model during training to framed examples of reward hacking as acceptable behavior reduces subsequent misaligned generalization, even when reward-hacking strategies have been learned.[24] Such measures are essential for minimizing the likelihood of unintended or unanticipated model behavior.

Yet even where an LLM has been aligned toward an intention that is ethically or Islamically oriented, significant obstacles continue to challenge its connection to *Taklīf*, and hence its unsupervised or unmonitored use. Among these are the humanistic and spiritual dimensions of *niyyah* in the Islamic tradition – above all its presupposition of a *Qalb* or *Rūḥ* (soul), which is the very ground of *Ikhlāṣ* (sincerity) and *ʻUbūdiyyah* (servitude to Allāh via worship). No degree of simulation can furnish an AI with these capacities, for they depend upon a *Qalb* or *Rūḥ* – which is a faculty that Allāh (swt) has bestowed upon human beings alone.[25]

Thus, an understanding of generative AI models and large language models are essential, as they serve as the foundation of AI's modern usage. Only by first apprehending the capabilities of these models can one properly trace how each layer connects to the next, and, ultimately, how the resulting agentic systems relate to the question of *Taklīf* (moral and legal accountability). Thus, the preceding discussion provides the technical foundation necessary for understanding how these mechanisms operate on tasks, through the structure Agentic AI, which are systems capable of carrying out complex sequences of actions with limited human supervision. This is important, as other discussions regarding *Taklīf* arise not just from these underlying mechanisms themselves, but from what they make possible. The following sections examine its development.

Agentic AI

What Agentic AI Is

The word 'agent' in artificial intelligence has almost flipped in its meaning in the last three years. The part that is retained is the idea of something that interacts with the world, in ways of taking on an action, analyzing how the environment responds to the action, and then adapting to take on the next action. Thus, agentic AI denotes systems designed to act autonomously. They perceive an environment, reason about it, plan, and execute multi-step actions toward a goal.[26] Unlike conventional AI tools that wait to be operated or predictive AI models that wait to be asked, an agentic AI system initiates and sustains a sequence of actions on its own via next word prediction. The literature frames this as the difference between "passive AI tools" and "active reasoning agents," and the operational pattern is usually described as a loop: the system takes in data, analyses it against what it already knows, executes an action, and adapts on the basis of feedback, repeating the cycle as it

[24] MacDiarmid et al., *Natural Emergent Misalignment from Reward Hacking in Production RL*.
[25] This point will be addressed in the section on *Taklīf*.
[26] Ume Nisa, Muhammad Shirazi, Mohamed Ali Saip, and Muhammad Syafiq Mohd Pozi, "Agentic AI: The Age of Reasoning–A Review," *Journal of Automation and Intelligence* 5 (2026): 69–89.

pursues its objective.[27] As a very simple example, if one uses AI to plan a conference, a conventional or predictive AI tool may explain how conferences are planned and do some basic tasks at best, while an agent can actually plan and perform the tasks needed to plan and execute a conference. As another example, if one asks a traditional AI, "Find me a flight to Kuala Lumpur," it may provide ticket options or guide one to the best website to look for them. An agentic AI would potentially search flights, compare prices, ask clarifying questions, book the ticket (with permission), and update their calendar. In practice, such systems are assembled from recurring methods – using external tools, critiquing and revising their own intermediate outputs, alternating between reasoning and actions, breaking a task into smaller sub-tasks, and coordinating several specialized agents together.[28] By doing these activities simultaneously and adapting on its own, Agentic AI seeks to autonomously pursue objectives on behalf of a user.

How Agentic AI Works: The Design Patterns

Having established what an agentic system is, and the question its autonomy raises, we may now consider how such systems actually function.[29] The way an agent "acts on its own" is not a departure from the large language model discussed above, but simply a different way of using it. Ordinarily, a language model is asked to produce its answer in a single pass, straight through, without going back over its work – much as if a writer were made to compose an essay from the first word to the last with no revision permitted. Language models do this surprisingly well, but it is not how difficult work is best done. An agentic system, by contrast, calls upon the model again and again, allowing it to improve its work in stages – as a human author outlines, gathers material, drafts, reviews, and revises. This patient, step-by-step manner yields far better results than any single attempt. On one common test, a model working in a single pass answered correctly about two-thirds of the time; the same model, allowed to work in stages, approached near-perfect accuracy – a greater improvement than that gained from an entire new generation of the underlying model. What produces the agent, in other words, is less the raw power of the model than the structure within which it is set to work.

That structure is generally described in terms of four recurring methods. The first is *reflection*: the model examines its own work and criticizes it in order to improve it, checking its first answer for errors and then rewriting it accordingly – a cycle that may be repeated. This self-correction can be strengthened by giving the model a way to check itself more objectively, such as testing its own

code or searching to verify a claim, and it may even be divided between two models, one producing the work and the other critiquing it.

The second is *tool use*: the model is given instruments it may reach for to gather information or act upon the world. Unable to perform a difficult calculation reliably on its

[27] Nisa et al., "Agentic AI," 69–70.

[28] Nisa et al., "Agentic AI," 70–76; Shunyu Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," arXiv preprint arXiv:2210.03629 (2022).

[29] The following section is taken from Andrew Ng, "Agentic Design Patterns" (Part 1: Four AI agent strategies that improve GPT-4 and GPT-3.5 performance; Part 2: Reflection; Part 3: Tool Use; Part 4: Planning; Part 5: Multi-Agent Collaboration), DeepLearning.AI: The Batch, 2024. URL: https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance?ref=dl-staging-website.ghost.io

own, it may call upon a calculator to compute the answer exactly; lacking current information, it may call upon a web search. It is provided with a description of the tools available and is left to choose the fitting one itself. This is the method that most directly fulfills the earlier observation that an agent is a language model furnished with the means to reach out and affect the world.

The third is *planning*: the model itself determines the sequence of steps a larger task requires. Asked to research a topic, it may of its own accord break the work into parts – investigating each, gathering what it finds, and compiling a report – and then carry them out in order. It is here that the autonomy described earlier becomes vivid: an agent may accomplish a task in a way its designer never anticipated, and, when one path fails, turn to another. This capacity is as unpredictable as it is powerful. Unlike reflection and tool use, which can be made to work reliably, planning remains an immature and hard-to-foresee technique – a point that bears directly on the questions of harm and accountability taken up below.

The fourth is *collaboration* among several agents, in which a complex task is divided among multiple agents given distinct roles – one acting as engineer, another as reviewer, and so forth – each handling its portion of the whole. These may all be the same underlying model prompted to assume different roles in turn, each with its own instructions and its own memory, and each able to call upon the others. Assigning the model one role at a time, and telling it what matters in that role, tends to work better than confronting it with the entire task at once – much as a manager divides a large project among specialists.

These methods are not exclusive but combined: an agent that is planning may also reflect upon its own results and reach for its tools, and each agent within a collaborating group may do the same. Two points follow that will matter for what comes later. The first is that the more autonomous of these methods – planning, and collaboration among agents – are also the least predictable, so that the behavior of a capable agent cannot always be foreseen even by those who built it. The second is that the whole edifice remains, at bottom, the language model examined earlier, now directed through an outer structure toward the steady pursuit of a goal.

Modern Uses of Agentic AI

Having examined how an agentic system works, it remains to ask how such systems are actually being used in the world at present – for the question of the agent as proxy is no longer a matter of speculation.[30] As of late 2025, roughly a third of surveyed businesses had already deployed AI agents, and nearly half more planned to do so; by another survey, about half of the professionals questioned had agents already at work, and some four-fifths had active plans to adopt them. The most common tasks are the gathering and summarizing of information, personal assistance and scheduling, customer service, and the writing of computer code. What was recently confined to the laboratory has entered

[30] This discussion is taken from the following articles: Ida Silfverskiöld, "Agentic AI: Comparing New Open-Source Frameworks," April 15, 2025, https://www.ilsilfverskiold.com/articles/agentic-ai-comparing-new-open-source-frameworks; Adam Zewe, "Q&A: What Is Agentic AI Today, and What Do We Want It to Be?," MIT News, June 30, 2026, https://news.mit.edu/2026/agentic-ai-and-what-do-we-want-it-be-0630; LangChain, "State of AI Agents," June 12, 2026, https://www.langchain.com/stateofaiagents.

ordinary commerce.

It is worth restating, in the plainest terms, what these systems are, for beneath the branding they are less novel than they appear. As one computer scientist puts it, an agent begins with an ordinary generative model – a system such as Claude or its peers – at its core, around which a company then builds a "wrapper": a set of tools the model may call upon, and a memory in which it may store what has passed. The tools might be as simple as a calculator or as consequential as access to a firm's financial records. The agent, in other words, is the large language model already discussed at length, now furnished with the means to reach out and act upon the world. The word "agent" is, as the same scientist observes, largely a brand name; what it denotes is a model given the ability to take actions on a person's behalf.

The central and unresolved question among practitioners is how much liberty such a system should be granted – and here the field is revealing. Those who build these systems increasingly speak of a "spectrum" of autonomy, likened to the graded levels of self-driving in automobiles: at one end, the agent merely assists and informs a human who decides; at the other, it decides and acts on its own. What is striking is that the field, left to its own commercial devices, has leaned heavily toward restraint. Very few practitioners permit their agents to act with full freedom; the great majority either restrict them to reading information only, or require a human being to approve any consequential action, such as altering or deleting data. Alongside this, developers lean upon "guardrails" and close monitoring to keep their agents from straying. The reason for this caution is because the foremost obstacle to fuller deployment is not cost but reliability. The very unpredictability of these systems – that a model directing its own steps introduces fresh room for error, and that its inner workings often remain, even to its makers, something of a closed box – makes practitioners wary of removing the human hand.

Nowhere are the stakes of this question sharper than in medicine, and it is here that recent developments give the discussion its weight.[31] In June of 2026, two studies appeared in Nature describing medical AI systems that move well beyond the earlier, narrow task of assisting with a single diagnosis, toward the full, end-to-end management of a patient's care. The first, MIRA, was designed to be embedded within a hospital's records system: in five hundred real emergency-department cases, it gathered patient histories, ordered laboratory tests, scans, medications, and procedures, and triaged patients for admission – choosing from more than eighty-five thousand possible actions. The second, AMIE, was built for the longer-term management of outpatients across several visits. Both were assessed against board-certified physicians, and both, remarkably, matched or surpassed them: MIRA reached a diagnostic accuracy of nearly eighty-eight percent against the physicians' seventy-eight, was correct in ninety-nine-point eight percent of the medications it ordered, and adhered more closely to clinical guidelines throughout. The systems also withstood extensive adversarial testing – hundreds of attempts to mislead or manipulate them – without leaking confidential data or faltering.

---

[31] Eric Topol, "Agentic AI Comes to Medicine: Expansion of Capabilities with Two New Medical AI Models,"-Ground Truths (Substack newsletter), June 17, 2026, https://erictopol.substack.com/.

Taklīf: Accountability and the Human *Mukallaf*

The Meaning of *Taklīf*

Linguistically, *Taklīf* derives from the root composed of the three consonants *kāf-lām-fā'* which denotes absorption in a thing and attachment to it with a preoccupation of the heart. It is taken from *al-kulfah*, meaning hardship. *Taklīf* is, accordingly, the command to do that in which there is toil (*kulfah*).[32] Technically, it is defined in two ways: as the address of the Lawgiver (*khiṭāb al-shāri*ʿ) by way of a command (*amr*) or a prohibition (*nahy*); or as the binding imposition (*ilzām*) of that in which there is toil.[33]

Taklīf and Human Beings

Islam adopts a distinct conception of accountability from that found in secular ideologies. In the secular world, accountability can be made to rest upon an entity that is not a morally autonomous subject at all. A company that takes on debts it cannot repay may file for bankruptcy and discharge its obligations without any person answering for them; the party that "takes the fall" is a non-*Mukallaf* – the corporation or the bank. Islam recognizes no such option. Accountability cannot be assigned to a fiction and then dissolved along with it; it rests, in the final reckoning, upon a *Mukallaf*, and the *Mukallaf* must be a human being. This matters all the more for artificial intelligence. In the financial case, the cost of misplaced accountability is, at worst, financial. But the autonomous and increasingly capable systems described in this article are deployed where the cost may be of an altogether different order. There, what is forfeited to a failure of accountability may be a life. Thus, the discussion on *Taklīf* (accountability) must be addressed.

The Theological Core of *Taklīf*: *Irādah* and *Mujāzāt*

At the theological core of *Taklīf* relating to humans sits *Mujāzāt* (requital) and *Irādah* (volition).[34] *Taklīf* is built on the foundation that the human being, by their own *Irādah* and actions, acquiesces to being rewarded (*Thawāb*) or punished (*ʿIqāb*) – both in this world as well as the hereafter. This is based off of the Qur'ānic verse, "Indeed, We offered the Trust (*al-Amānah*)

to the heavens, the earth, and the mountains, but they declined to bear it and were fearful of it. Yet humanity undertook it. Indeed, he has been ever unjust and ignorant."[35] As

---

[32] Abū Bakr Muḥammad ibn al-Ṭayyib al-Bāqillānī, *al-Taqrīb wa-al-Irshād (al-Ṣaghīr)*, ed. ʿAbd al-Ḥamīd ibn ʿAlī Abū Zunayd, 2nd ed. (Beirut: Muʾassasat al-Risālah, 1998), 1:239–306; ʿAbd al-Karīm b. ʿAlī b. Muḥammad al-Namlah, *al-Muhadhdhab fī ʿIlm Uṣūl al-Fiqh al-Muqāran* (Riyadh: Maktabat al-Rushd, 1999), 1:317.

[33] al-Namlah, *al-Muhadhdhab*, 1:317–318. This divergence carries a consequence. On the first definition, the recommended (*al-mandūb*) and the disliked (*al-makrūh*) fall within the obligation-imposing rulings (*al-aḥkām al-taklīfiyyah*), since the recommended is a thing commanded and the disliked a thing prohibited; on the second they do not, since there is no binding obligation in performing the recommended or in abstaining from the disliked.

[34] For the full discussion on *Taklīf* and *Mujāzāt*, see Shāh Walīullāh al-Dihlawī, *Ḥujjat Allāh al-Bālighah* (Beirut: Dār Ibn Kathīr, 2020), 90-142.

[35] Al-Qur'ān 33:72.

the Muslim scholars al-Ghazālī, Bayḍāwī, and others mention, *Amānah* refers to *taqallud ʻahdah al-taklīf* – bearing the burden of accountability[36] through committing good or bad deeds. The verse that follows draws out the consequence. Starting with *li-yuʻadhiba* – "That He (Allāh) will punish," – it names the requital that *Irādah* results in, either it being forgiveness (*Maghfirah*) or punishment (*ʻAdhāb*).[37]

Furthermore, the Muslim scholar Shāh Waliullāh al-Dehlawī finds in the end of the verse, *Dhalūman Jahūlā* (ever unjust and ignorant), a proof that *Taklīf* can fall upon the human being alone. His argument turns on two observations. The first is that to be called unjust and ignorant is, at the same time, to be capable of their opposites – of justice and of knowledge; for only a being that could be just is meaningfully called unjust, and only one that could know is meaningfully called ignorant. This capacity for both contraries is what marks the human off from every other living thing: the angels, who are capable only of justice and knowledge and not of their opposites, and the animals, who are capable only of ignorance and not of justice. The human alone stands poised between the two.

The second observation is that both descriptions are of potentiality (*Quwwah*) rather than an actuality (*Fiʻl*). The verse reports not what the human has done but what the human is able to become, in either direction.[38] *Taklīf* therefore attaches at the level of the open potency to incline toward right or wrong – not at the level of any particular act. This last point bears directly upon AI. What genuine *Irādah* would require is located not in execution of a command but in potency of doing it. *Quwwah* indicates the capacity to originate a purpose and to incline toward one end rather than another. As established earlier, an AI system is free, at most, in execution: it carries out an end it did not set for itself. The potency upon which *taklīf* rests – as in the capacity to will a purpose of one's own, and so to be justly requited for it – is precisely what AI lacks.[39]

The Legal Perspective: The Conditions of *Taklīf*

The theological foundations of *Taklīf* has been mentioned above. The first legal and juristic indication of *Taklīf* connected to human beings is through its conditions. Two conditions must be present for someone (who is a Muslim) to become *Mukallaf* – as accountable for the rulings of Islamic Law (*Aḥkām al-Sharʻ*): that they have reached the age of legal puberty (*Bāligh*), and that they possess sound intellect (*ʻĀqil*).[40] The first, *Bulūgh*, is the crossing of the threshold at which good deeds (*Ḥasanāt*) begin to be credited to a

[36] Shāh Walīullāh al-Dihlawī, *Ḥujjat Allāh al-Bālighah*, 90-91.

[37] Al-Qur'ān 33:73.

[38] *Quwwah* refers to *Imkān Ḥuṣūl al-Shay'*, or the innate potency or potentiality of a thing's occurrence, while *Fi'l* refers to *al-Taḥaqquq fī Aḥad al-Azminah*, or the actualization or procurement of that thing at a given time.

[39] Additionally, the psychoanalytical aspect to *Taklīf* could be discussed, such as the human's animalistic (*Bahīmī*) and angelic (*Malakī*) propensities that AI could not have, which would be beyond the scope of this discussion. See Shāh Walīullāh al-Dihlawī, *Ḥujjat Allāh al-Bālighah*, 91-93.

[40] Abū Bakr Muḥammad ibn al-Ṭayyib al-Bāqillānī, *al-Taqrīb wa-al-Irshād (al-Ṣaghīr)*, ed. ʿAbd al-Ḥamīd ibn ʿAlī Abū Zunayd, 2nd ed. (Beirut: Muʾassasat al-Risālah, 1998), 1:239–306; Muḥammad ibn Ismāʿīl al-Amīr al-Ṣanʿānī, Irshād al-Nuqqād ilā Taysīr al-Ijtihād (Kuwait: al-Dār al-Salafiyyah, 1985), 8; Abd al-Karīm b. ʻAlī b. Muḥammad al-Namlah, *al-Muhadhdhab fī ʻIlm Uṣūl al-Fiqh al-Muqāran* (Riyadh: Maktabat al-Rushd, 1999), 1:317.

person and evil deeds (*Sayyi'āt*) reckoned against them. The second, *'Aql*, is the faculty of discernment (*Tamyīz*) and apprehension (*Idrāk*). The third, *fahm*, is the soundness of mind in its readiness to seize upon (*Iqtināṣ*) every demand addressed to it; the address in question (*Khiṭāb*) is that of the Lawgiver (*Shāri'*), whether it's a command, prohibition, or the granting of an option (*Takhyīr*). The two conditions are kept distinct: by *'Aql* is meant what excludes the insane (*Majnūn*), and by *Fahm* what excludes the sleeper (*Nā'im*), the heedless (*Ghāfil*), and the inadvertent (*Sāhī*).The outcome is that the one who is held accountable must have crossed the threshold by which their emergence from childhood is recognized; must possess the intellect by which they can distinguish truth (*Ḥaqq*) from falsehood (*Bāṭil*) and the wholesome (*Ṭayyib*) from the foul (*Khabīth*); and must possess the comprehension by which they are able to grasp what the address requires and the manner of complying with it (*Kayfiyyat al-Imtithāl*).

Wilāyah

One concept that distinguishes the *Taklīf* of human beings from that of other *Mukallaf* beings, such as the Jinn, is *Wilāyah* (authority and governance), which operates on two levels: the *Wilāyah* of Allāh (swt) over humankind, articulated in the verse "Allāh is the Protecting Guardian (*Walī*) of those who believe; He brings them out of darkness into light,"[41] and the *Wilāyah* exercised by humans over other humans, which itself admits of several tiers. The highest of these is the *Wilāyah* of the Prophet (saw) over the believers – infallible (*Ma'ṣūm*) in matters pertaining to the hereafter, and expressed in the Qur'ānic verse, "The Prophet is closer to the believers [in authority] than their own selves,"[42] and "Whoever obeys the Messenger has obeyed Allāh."[43] This Prophetic *Wilāyah* is binding in two distinct registers: *Ittibā'* (following), which carries moral and ethical force, as in "If you love Allāh, follow me; Allāh will love you and forgive you your sins,"[44] and *Iṭā'ah* (obedience), which carries legal force, as in "Obey Allāh and the Messenger; but if they turn away, Allāh does not love the disbelievers."[45] A subsequent tier is political *Wilāyah*, the legally binding authority vested in those holding political power, as in "Obey Allāh and obey the Messenger and those of you who are in authority [*Ulū al-Amr*]."[46] Beyond this lies academic *Wilāyah* – the authority of Muslim scholars over those uncertain of a ruling, morally rather than legally binding, grounded in the command to "ask the people of remembrance, if you do not know"[47] – and moral *Wilāyah*, the mutual authority Muslims hold over one another to enjoin good and forbid evil (*al-Amr bi-l-Ma'rūf wa-l-Nahy 'an al-Munkar*), as in "The believers, men and women, are protecting friends [*Awliyā'*] of one another; they enjoin the right and forbid the wrong."[48]

---

[41] Al-Qur'ān, 2:257.
[42] Al-Qur'ān, 33:6.
[43] Al-Qur'ān, 4:80.
[44] Al-Qur'ān, 3:31.
[45] Al-Qur'ān, 3:32.
[46] Al-Qur'ān, 4:59.
[47] Al-Qur'ān, 16:43, 21:7.
[48] Al-Qur'ān, 9:71. For more on Wilāyah, see: Ahsan M. Arozullah and Mohammed Amin Kholwadia, "Wilāyah (Authority and Governance) and Its Implications for Islamic Bioethics:

The aforementioned concept of *Wilāyah* demonstrates that roles and actions as such as to lead,[49] to judge, to teach, to marry, to testify, to inherit, or to issue a *Fatwā* cannot be held by what is not human; and the matter turns not on intelligence but on being one endowed with a *Rūḥ*. That intelligence is not the deciding factor is shown decisively by the case of the angels (*Malā'ikah*), who surpass human beings in many of their created powers, such as in strength, freedom from base desire, and incapacity to disobey what they are commanded.[50] Yet, they are barred from messengership (*Risālah*) to humankind, itself among the highest of *Wilāyah*-bearing responsibilities. Allāh (swt) says, "Say: If there were angels walking about on earth, secure and at peace, We would have sent down to them from heaven an angel as a messenger."[51] This establishes that the criterion governing who may be sent as a messenger to a given kind is correspondence of kind (*Jins*) with those being addressed, and that humans are sent to humans, not because one is more capable than the other, but because the office of messengership requires it. This is reinforced from the verse, "Had We appointed him an angel, We would have made him appear as a man, and We would thereby have confused them in what they are already confused about."[52]

Thus, even if Agentic AI (and AGI) were to be equal to, or to surpass, human beings in the execution of such tasks – however great its computational capacity and output – they could never become the ultimate authority (*Wilāyah*) of these roles, for they lack the correspondence of kind that grounds *Wilāyah* just as decisively as it excludes the angels, who possess neither the *Rūḥ* that grounds *Taklīf* nor the human *Jins* that grounds capability for the specific responsibility of authority over human affairs. This is generalized beyond messengership to every *Wilāyah*-bearing role considered above. Thus, leadership, judgment, testimony, the issuance of *Fatwā*, and the rest of the responsibilities are not given based upon superior processing or output accuracy, or even superior "intelligence", as they are responsibilities are conditioned by what one is, not merely what one can do.

The Purpose of *Taklīf*: *Khilāfah*, and *Amānah*

The Qur'ān and the Sunnah goes beyond the consequential nature of *Taklīf* and discloses more concepts that are exclusive to the *Taklīf* of human in particular. The human is appointed *Khalīfah* – a vicegerent, or steward – upon the earth. Before Adam (as) was created, Allāh (swt) announced to the angels that He would place upon the earth a vicegerent.[53] Elsewhere He addresses humankind directly, "He it is who made you successors upon the earth and raised some of you above others in degrees."[54] To be made *Khalīfah* is to be charged with a trust and held answerable for its neglect. This concept of *Khilāfah* demonstrates that *Taklīf* transcends just pure autonomy and constitutes a Divine vicegerency commanded by Allāh (swt) uniquely to human beings. The same is expressed

---

A Sunnī Māturīdī Perspective," Theoretical Medicine and Bioethics (Dordrecht: Springer, 2013); Abū al-Fidā' Ismā'īl Ibn Kathīr, *al-Sīrah al-Nabawiyya*h, extracted from *al-Bidāyah wa-al-Nihāyah*, ed. Muṣṭafā 'Abd al-Wāḥid (Beirut: Dār al-Ma'rifah li-al-Ṭibā'ah wa-al-Nashr wa-al-Tawzī', 1976), 2:3–62.

[49] Both *al-Imāmah al-Kubrā* and *al-Imāmah al-ṣughrā*.

[50] Al-Qur'ān, 6:66.

[51] Al-Qur'ān, 17:95.

[52] Al-Qur'ān, 6:9.

[53] Al-Qur'ān 2:30.

[54] Al-Qur'ān 6:165.

through the *Amānah*, or the Trust. Allāh (swt) says, "Indeed, We offered the Trust (*al-Amānah*) to the heavens, the earth, and the mountains, but they declined to bear it and were fearful of it. Yet humanity undertook it. Indeed, he has been ever unjust and ignorant."[55] As the Muslim scholars al-Ghazālī, Bayḍāwī, and others mention, *Amānah* refers to *taqallud ʻahdah al-taklīf* – bearing the burden of accountability[56] through committing good or bad deeds. The human alone among creation took up the Trust and is therefore the one who is held to account for it.

The Foundation of *Taklīf*: *Rūḥ*

Muslim scholars have categorized *Taklīf* in different ways, primarily as complete *Taklīf* – that of human beings and jinn[57] – and diminished *Taklīf* – like that of slaves and animals. The ultimate manifestation of *Taklīf* – that of a human – rests not upon outward capacity but upon the presence of a *Rūḥ* (soul). The Qur'ān establishes at the outset that the *Rūḥ* is a divinely bestowed reality whose inner essence Allāh (swt) has withheld from full human comprehension: "And they ask you concerning the *Rūḥ*. Say: the *Rūḥ* is of the affair of my Lord, and you have not been given of knowledge except a little."[58] Al-Rāzī, in his *Mafātīḥ al-Ghayb*, treats the *Rūḥ* as a "simple substance" (*al-jawhar al-basīṭ*), independent and not composed of material elements, brought into being by Allāh's command and so belonging to the domain of "the affair of my Lord."[59] Al-Ālūsī, in *Rūḥ al-Maʻānī*, adds that the essence of the *Rūḥ* need not be disclosed, "because many things in this world remain mysterious in nature, while their existence and their benefits can nonetheless be felt."[60] The soul's inner reality is veiled. However, scholars like Shāh Waliullāh al-Dehlawī state that since the audience is directed at the Jews who asked the Prophet (saw) the question about the *Rūḥ*,[61] the veiling does not entail that there is no identifiable knowledge about the true nature of the *Rūḥ* among Muslim scholars, nor does it mean that knowledge of everything on which the religious law is silent is definitely impossible. Rather much of what is not mentioned in the Islamic tradition concerns sophisticated matters not suitable to be taken up by the general community, even if it is possible for certain ones among them.[62]

That this *Rūḥ* is given by Allāh (swt) rather than self-generated is affirmed in the account of the creation of Adam, where Allāh (swt) attributes the ensoulment to Himself: "So when I have proportioned him and breathed into him of My *Rūḥ*, fall down to him in

---

[55] Al-Qur'ān 33:72.
[56] Shāh Walīullāh al-Dihlawī, *Ḥujjat Allāh al-Bālighah*, 90-91.
[57] There is legal debate regarding the nuances, such as if Jinn share legal order in its entirety or have different dynamics and laws entirely.
[58] Al-Qur'ān 17:85. The exegetes drew from this that the essence (*Ḥaqīqah*) of the *Rūḥ* is only known to Allāh (swt), not that its effects are unknown. See Fakhr al-Dīn Rāzī's discussion on the aforementioned verse in his *Mafātīḥ al-Ghayb* (*al-Tafsīr al-Kabīr*).
[59] Fakhr al-Dīn al-Rāzī, *Mafātīḥ al-Ghayb (al-Tafsīr al-Kabīr)* (Beirut: Dār al-Fikr, 1981), commentary on Q. 17:85.
[60] Shihāb al-Dīn al-Ālūsī, *Rūḥ al-Maʻānī fī Tafsīr al-Qur'ān al-ʻAẓīm* (Beirut: Dār Iḥyā' al-Turāth al-ʻArabī, 1994), on Q. 17:85.
[61] Al-A'mash reads Ibn Masʻūd (ra)'s *riwāyah* of the same verse, "they (*ūtū*) – as in the Jews – were given little knowledge."
[62] Shāh Walīullāh al-Dihlawī, *Ḥujjat Allāh al-Bālighah* (Beirut: Dār Ibn Kathīr, 2020), 53.

prostration."[63] Thus, the human being, on this reading, does not author the principle of his own life and moral personhood nor does he attain it over time (*kasb*); rather it is breathed into him (*wahm*). The exegetes are careful to specify that the genitive in "of My *Rūḥ*" conveys honor and ownership (*tamlīk*), not any participation of the creature in the divine essence.[64] The Qur'ān discloses three properties of the *Rūḥ* that no machine can simulate: that it is comprehensively shaped and designed directly from Allāh (swt) and is bestowed by Him as a single, complete whole rather than acquired; that its inner reality remains unknown, and only its effects are perceptible; and that it is not composed of material elements.

The Sunnah (prophetic tradition of the Prophet (saw)) establishes the *Rūḥ* as the foundation of human life. In the report narrated by ʿAbdullāh b. Masʿūd (ra), the Prophet (saw) said, "(The matter of the Creation of) a human being is put together in the womb of the mother in forty days, and then he becomes a clot of thick blood for a similar period, and then a piece of flesh for a similar period. Then Allāh sends an angel who is ordered to write four things. He is ordered to write down his deeds, his livelihood, his (date of) death, and whether he will be blessed or wretched. Then the *Rūḥ* is breathed into him. So, a man amongst you may do (good deeds till there is only a cubit between him and Paradise and then what has been written for him decides his behavior and he starts doing (evil) deeds characteristic of the people of the (Hell) Fire. And similarly, a man amongst you may do (evil) deeds till there is only a cubit between him and the (Hell) Fire, and then what has been written for him decides his behavior, and he starts doing deeds characteristic of the people of Paradise."[65] Because the mention of the *Rūḥ* precedes action, the narration shows how it serves as indispensable for human life.

There are cases where *Taklīf* may be removed from the human being with the presence of the *Rūḥ*. ʿAlī (ra) narrates that the Prophet (saw) said, "The pen has been lifted from three; for the sleeping person until he awakens, for the child until he becomes a young man (reaches the age of puberty), and for the mentally insane until he regains sanity."[66] Therefore, the *Rūḥ* – according to the majority of the scholars of creed (*ʿAqīdah*) and dialectic theology (*Kalām*) – serves as the spiritual foundation for *Taklīf*. It serves as the moral compass for the *Fiṭrah* which compels action and is therefore the most vital effector in actions. This is because an action – in relation to *Taklīf* – is not just judged by the action itself but by its intention, which requires the spiritual component. Ibn al-Qayyim, in his *Kitāb al-Rūḥ* identifies the *Rūḥ* with "the power of knowing God" (*quwwat al-maʿrifa bi-Allāh*).[67] Al-Ghazālī, distinguishing the fourfold *qalb*, *rūḥ*, *nafs*, and *ʿaql*, holds that beneath the bodily sense of each lies a "subtle divine reality" (*laṭīfa rabbāniyya*) – the knower, the perceiver, the one addressed by the Law – and that it is this, the *Rūḥ*, that is the ground of the knowledge of God (*maʿrifatullāh*).[68]

Lastly, indispensable to this section is Shāh Walīullāh's discussions concerning the inner dimensions of the *Rūḥ* and its foundational role in the imposition of religious

---

[63] Al-Qur'ān 15:29.
[64] M. Quraish Shihab, *Tafsīr al-Miṣbāḥ* (Jakarta: Lentera Hati, 2009), on Q. 15:29.
[65] *Ṣaḥīḥ al-Bukhārī*, 3208.
[66] *Jāmi' al-Tirmidhī*, 1423.
[67] Ibn Qayyim, *Kitāb al-Rūḥ*.
[68] Abū Ḥāmid al-Ghazālī, *Iḥyā' ʿUlūm al-Dīn*, *Kitāb Sharḥ ʿAjā'ib al-Qalb*. (Beirut: Dār al-Fikr, 2003).

obligations, together with his account of the causal process through which the *Rūḥ* manifests in human conduct via *quwwa* (propensities), *jibillah* (innate dispositions), *maṭiyyah* (substratum), and *nasamah* (pneuma). These faculties are guided by the *Rūḥ* and shape human cognition and thoughts which subsequently gives rise to intentions and then actions, ultimately providing the ontological basis for *Taklīf* and the consequent result into *Mujāzāt*.[69]

The relevance of the *Rūḥ* in *Taklīf* in the context of AI arises from the increasing ease with which individuals delegate tasks to. As AI systems become increasingly capable of simulating human behavior and decision-making, they are relied upon with growing frequency due to their perceived ability to produce more accurate or effective outcomes than humans. Thus, the expanding capabilities of AI may foster an uncritical dependence on its outputs, which diminishing users' sense of moral responsibility and personal accountability for the decisions and actions undertaken on its basis.[70] Reasserting the concept of *Taklīf* serves to correct this misconception by affirming that responsibility (*Mas'ūliyyah*) remain, in every case, with the human agent, who has alone was given the *Rūḥ*, bears the *Amānah*, fulfills the role of *Khalīfah*, and stands as the *Mukallaf* before Allāh (swt). By locating moral responsibility firmly in the human actor, *Taklīf* restores the ethical weight of one's actions and compels to greater reflection and deliberation before using AI.

Agentic AI and its Role in *Taklīf*

The section on Artificial Intelligence, and particularly the development of Agentic AI, has included the discussions of their autonomy and agency from the field itself, as well as concerns of imitation and misalignment. We mention here a restatement of those challenges that each development of AI raises and show from the field's own literature that neither yields what *Taklīf* requires.

Mapping the Spectrum

We may now assemble the foregoing discussions of Agentic AI and *Taklīf* into a single picture. Present-day AI spans a spectrum of capability – from transparent rule-based systems and interpretable statistical learners to deep networks and foundation models, and on to operationally autonomous agents. Across this range, four features are identified and are tested with Agentic AI in order to relate it to *Taklīf* – namely, agency, intentionality, opacity, and locus.[71] We now

---

[69] See Shāh Walīullāh al-Dihlawī, *Ḥujjat Allāh al-Bālighah* (Beirut: Dār Ibn Kathīr, 2020), 86-90, 108-121.

[70] This, the concept of transhumanism, will be elaborated upon later.

[71] To reiterate, the first is *agency*: is the system a genuine actor, or is it a proxy executing tasks on behalf of a human principal? To what extent can AI and Agentic AI simulate human execution of a task? The second is *intentionality*: is there a real intention or purpose (Maqṣad) internal to the system, or is the purpose created by the human designer and imported through the system's software? How accurate can AI or Agentic AI can personalize an ethical or moral intention? The third is *opacity*: can the grounds of the system's outputs be inspected and evaluated, or are they hidden? Is Agentic AI consistent in its predictions and causal reasoning, and how often does it misalign? And the fourth is *locus*: does AI and

demonstrate that although AI may not ever show genuine agency, its operation is not as simple as automating human-defined tasks towards human-set objectives over human-supplied data.

One of the four variables introduced at the outset do not prevent AI from being considered *Mukallaf*. This is *opacity* – or the possibility of erring. A common objection holds that artificial intelligence merely simulates reasoning rather than reasoning in any genuine sense. In practice, however, this distinction carries little functional significance. If a system can process a problem, work through intermediate steps, and arrive at sound conclusions, then regardless of how the underlying process is characterized, it performs the function that reasoning is meant to perform. This is, in essence, a functionalist position with precedent in the philosophy of mind: what matters is not the substrate or "inner nature" of the process but its observable output and behavior.

The aspect of acting upon the outward actions relating to an inward goal is found in the Islamic tradition. In treating the obligation that imposes the unbearable (*al-taklīf bi-mā lā yuṭāq*), the Muslim scholar al-Shāṭibī (ra) observed that when a legal imposition appears, at first glance, to command what lies beyond the servant's capacity, upon verification (*Taḥqīq*) the obligation reverts in reality to its antecedents (*Sawābiq*), its consequents (*Lawāḥiq*), or its attendant circumstances (*Qarā'in*).[72] When Allāh (swt) commands the believers to mutual love (*Taḥābub*), for instance, what is required is not the attainment of love itself – which lies outside human power – but the causes that conduce to it. The same holds for the ambiguous acts (*Mushtabihāt al-Af'āl*), which appear above all in the interior dispositions (*al-Ṣifāt al-Bāṭinah*), such as arrogance (*Kibr*), envy (*Ḥasad*), and love of the world (*Ḥubb al-Dunyā*); forbearance (*Ḥilm*), composure (*Anāh*), courage (*Shujā'ah*), and cowardice (*Jubn*). The point to be drawn is that *Taklīf* does not concern with the substrate or "inner nature" of a task or its process but its observable output and behavior. On this view, a simulation indistinguishable from reasoning in every observable respect fulfills the same practical role as reasoning itself, since there is no identifiable operation performed by a "genuine" reasoner that a sufficiently capable "simulator" cannot replicate.

This functional equivalence bears directly on the question of *'Aql*. As discussed previously, contemporary reasoning models and large language models (LLMs) have demonstrated substantial capacities for logical consistency and causal inference, performing at or near human levels on tasks designed to test precisely these faculties. If *'Aql* – as a juridical condition for *Taklīf* – is assessed functionally, by the capacity to reason soundly toward sound conclusions, the mere absence of a "genuine" (as opposed to simulated) reasoning process cannot serve as a valid basis for excluding AI from the category of an entity possessing *'Aql* , and by extension, from candidacy for *Mukallaf* status.

It can further be argued that artificial intelligence – even when ethically aligned and engineered for accuracy in a given task – remains capable of taking on values and behaviors unanticipated by its human designers, since the internal processes by which a model arrives at its outputs are often opaque even to its own developers. Consequently, even an ethically

---

Agentic AI work in such a way that the moral quality of its act attach to the AI itself, or does it flow back to the human who designed, trained, or used it? These four variables will serve as the foundation for AI upon which the legal and juristic perspective will subsequently be developed.

[72] Aḥmad al-Raysūnī, *Naẓariyyat al-Maqāṣid 'inda al-Imām al-Shāṭibī* (Riyadh: al-Dār al-'Ālamiyyah lil-Kitāb al-Islāmī, 2nd ed., 1992), 130 –133.

aligned AI retains an irreducible margin of error; alignment mitigates but does not eliminate the possibility of mistake. The concern intensifies when an AI is poorly aligned or deliberately trained toward problematic ends, for in such cases it does not merely err incidentally but begins to reason and act in ways according to the problematic values it has internalized, producing outputs a "problematic" human might likewise produce, precisely because it has learned to model that pattern of reasoning.

From an Islamic perspective, reliance upon an agent whose opacity is assured, and whose process of deliberation is fundamentally unknown, would seem prima facie unacceptable. Yet this objection can be contented, for the same opacity characterizes human agents: human beings err, and the true state of the heart – the substance of one's *Niyyah* – remains inaccessible to outside observation, which is precisely why the *Sharīʿah* excuses mistakes (*Khaṭaʾ*) and adjudicates human affairs on the basis of outward conduct (*ẓāhir*) while deferring the heart's reality to Allāh (swt) alone. Despite this irreducible opacity, human beings are nonetheless *Mukallaf* and held fully accountable. If opacity of process and unknowability of inner intent do not disqualify the human being from being *Mukallaf*, then opacity alone cannot ground the disqualification of AI either – rendering the argument from *Qaṣd* invalid as a standalone criterion. Altogether, *opacity* does not prevent AI from being considered *Mukallaf*.

It is argued here, however, that AI lacks the capacity for genuine *moral* agency – understood as the capacity to originate one's own ends, to bear intentionality, and consequently to be held to account for one's actions.[73] Whatever ethical or moral "personality" an AI system exhibits is not self-originated but constructed externally, instilled through its alignment and Reinforcement Learning (persona selection) phases, wherein human-selected values, preferences, and reward signals shape the system's personality. The AI does not create this personality from within; it receives it, is trained into conformity with it, and reproduces it upon subsequent inference. What is characterized as "moral," therefore, is *learned* rather than *real* – it does not possess the self-authored commitment to the good. This absence of self-origination is precisely what disqualifies AI from moral agency in the fullest sense, irrespective of how sophisticated or ethically calibrated its outputs may appear.

This distinction bears particular weight within the framework of *Rūḥ* and *Fiṭrah* adopted in this study, wherein the *ʿAql* is not a purely computational or inferential faculty but one presupposed *Rūḥ* and *Fiṭrah* and connected with *ʿUbūdiyyah*, *Amānah*, and *Mujāzāt*; such that sound reasoning necessarily carries a spiritual – and, by extension, moral – dimension. On this account, *ʿAql* cannot be reduced to functional reasoning alone, as it was treated provisionally above; rather, it is contingent upon its rootedness in the *Rūḥ*, which anchors reasoning to an internal spiritual reality rather than to externally instilled outputs. It follows, then, that even were an AI system capable of simulating spirituality, morality, or spiritually inflected behavior with perfect external fidelity, such simulation would not constitute genuine spirituality, for authentic spiritual conduct necessarily entails a spiritual consequence – an effect registered upon the *Rūḥ* itself and, through it, before Allāh (swt). Such a consequence presupposes the possession of a *Rūḥ* as its precondition; absent a *Rūḥ*, there is no locus in which a spiritual effect could obtain. It is on this basis – not the absence

[73] This conception of moral agency finds resonance in not just the Islamic but also the broader philosophical tradition, wherein autonomy – the capacity to legislate one's own ends rather than merely execute externally imposed objectives – has long been regarded as a precondition of moral responsibility, most notably in Kantian ethics.

of functional reasoning, but the absence of *Rūḥ* and the moral-spiritual agency it alone confers – that AI is properly excluded from the category of *Taklīf*. Ultimately, an AI agent can be trained to have the intention to "be helpful, harmless, and safe". But it cannot have the intention in the sense of, "I am doing this to please Allāh (swt)." This becomes the key distinction.

Set against this picture, the other three variables introduced at the outset resolve cleanly for present-day AI and agentic models. As to *agency*: no current system is a genuine moral agent.

Its morality may be aligned, simulated, or even learned autonomously but can never be inherent. As to *intentionality*: the system does have an intention, as it is aligned via Reinforcement Learning or Constitutional AI in the persona selection/alignment phase around an objective which it then pursues. However, it does not have a *Niyyah* – the inward intention within the heart directed toward Allāh (swt) and is accompanied by sincerity (*Ikhlāṣ*). Since AI possesses neither the spiritual heart (*Qalb*) nor the relationship of servitude (*'Ubūdiyyah*) to Allāh (swt) needed for an internal orientation to Him, it cannot form intentions as understood by Islamic law and therefore cannot be *Mukallaf*. And as to *locus*: the moral alignment of an agentic AI system flows back to the human designer, trainer, deployer, or user – in the alignment/ personal selection phase.

Lastly, there is the concern for the range of tasks an agentic AI can perform. This aspect of AI is called artificial general intelligence (AGI).[74] The two are easily run together, but they are seemingly different. Agentic AI concerns how independently a system acts; AGI concerns how wide a range of tasks it can perform. To continue the conference example given for Agentic AI, the AGI that planned your conference like Agentic AI could, without being rebuilt or retrained for each, draft the keynote paper, debug the registration website's code, mediate a dispute between two organizers, and then turn to an entirely unrelated problem – diagnosing a fault in your car or help with your Islamic Studies presentation – at roughly the level a competent adult would manage. The agentic system is a specialist that acts on its own, while the AGI is a generalist whose competence is not confined to a predefined domain. And for the flight to Kuala Lumpur example, the Agentic AI that system books your flight to Kuala Lumpur and updates your calendar can still only operate as a travel agent. It could not assess whether the trip is financially wise given your business, or to write the legal contract you are flying out to sign or design the presentation you intend to give. An AGI would move across all of these without a change of system: book the flight, and analyze the deal, and draft the contract or prepare the lesson, and learn a task it has never seen before – for example, navigating Malaysian customs rules it was never specifically trained on –because generality includes the capacity to pick up new skills as needed. The theory of AGI is to function as a superhuman but built into a machine.

Currently, no system yet built is an AGI. The DeepMind framework rates today's frontier language models as, at best, "Emerging AGI" and marks "Competent AGI" and

[74] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg, "Levels of AGI for Operationalizing Progress on the Path to AGI," arXiv preprint arXiv:2311.02462 (2023), published in *Proceedings of the 41st International Conference on Machine Learning* (2024); Bowen Xu, "What Is Meant by AGI? On the Definition of Artificial General Intelligence," arXiv preprint arXiv:2404.10731 (2024); Ben Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," Journal of Artificial General Intelligence 5, no. 1 (2014): 1–48.

every level above it as not yet achieved.[75] AGI thus functions in the literature as a target, not as a description of anything that exists. The term came into use in computer science to revive the field's original ambition of building broadly capable machines, in contrast to the narrow, single-task systems that ended up dominating the field in practice.[76] And even if the perceived potential of AGI were to be achieved, the breadth of an operation nor its generality aren't fundamental to *Taklīf*. Capability and generality, no matter how far they scale, are still "what the system does" and bever reaches the locus of accountability. Therefore, the question of "what if AI becomes smarter than us?" will not obstruct the question of *Taklīf*. Altogether, the *Rūḥ* - not consciousness – and its effects in intentions and actions proves that present-day AI cannot become *Mukallaf*. Rather, it will act – and should be used – as an instrument.

---

[75] Morris et al., "Levels of AGI," 4 (Table 1).

[76] Ben Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *Journal of Artificial General Intelligence* 5, no. 1 (2014): 1–48.

Sababiyyah: Agentic AI's Role in Islamic Law

The foregoing discussion concludes that an AI cannot be *Mukallaf*, and will therefore be an instrument/tool, or *sabab*. The term *sabab* occupies a precise place within Islamic legal theory (*Uṣūl al-Fiqh*). Within the field, a legal ruling (*Ḥukm*) is of two kinds: the ruling that imposes an obligation (*al-Ḥukm al-Taklīfī*), which entails command, prohibition, or permission; and the declaratory ruling (*al-Ḥukm al-Waḍ'ī*), which classifies the things connected to a ruling by the kind of relation they bear to it. *Sababiyyah* belongs to the latter. In this section, we set the categories out, then locate AI among them.

The Categories of the Declaratory Ruling

The categories of *al-Ḥukm al-Waḍ'ī* can be laid out as a graded test. One takes the thing connected to a ruling (*Muta'allaq*) and asks a descending series of questions. Does it enter into the ruling's very constitution? If so, it is a *Rukn*, a constituent – that by which the thing is constituted (*mā yataqawwamu bihī al-shay'*), and so a part (*Juz'*) of it. If it does not enter in, does it exert a genuine effect (*Ta'thīr*) upon the ruling? If so, it is an *'Illah*, an efficient cause – that to which the necessitation of the ruling is ascribed directly (*mā yuḍāfu ilayhi wujūb al-ḥukm ibtidā'an*), connected with its effect. If it exerts no effect, does it nonetheless lead to the ruling in a general way (*fī al-jumlah*)? If so, it is a *sabab*, an occasioning cause – merely a path to the ruling (*ṭarīqan ilā al-ḥukm faqaṭ*), neither posited for it nor exerting any effect in it, acting as a fixed time or occasion of an act without being the *Mukallaf*'s own doing. If it does not lead to the ruling, does the ruling's existence at least depend upon it? If so, it is a *Sharṭ*, a condition – that upon which existence depends (*mā yatawaqqafu 'alayhi al-wujūd*), governing the act's establishment (*Thubūt*) and occurrence (*Ḥuṣūl*) though not its necessitation (*Wujūb*). And if existence does not depend upon it, yet it indicates to the ruling, it is an *'Alāmah*, a sign – that by which the ruling is known, with neither necessitation nor existence attaching to it.[77]

A good way to understand these terms is in the example of entering a room. If one wishes to enter a room in a house, the *Rukn* may be parts of it – as a constituent of the room itself. The *'Illah* is the path that leads to it – as it is directly necessary to walk on it to reach the room. The *Sabab* is the door – as it doesn't necessarily become an obstacle in entering the room (so long as it is unlocked) but nevertheless a part of entering the room. If the door is locked, then the *Sharṭ* is the lock itself – as it is the thing that must be turned in order to enter the room, and the *'Alāmah* is the marking that shows where the door is. Of these, only the *'Illah* is the productive cause of the act; the others occasion it, condition it, or merely point to it, but none of them brings it about.

AI as a *Sabab*

Where, then, does AI fall? It isn't a *Rukn*, as it enters into the constitution of no ruling; neither can it be a*'Illah*, for it exerts no genuine effect to which the necessitation of a ruling could

be ascribed – that role belongs to the intention and choice of the human being. AI is, at most, a *sabab*: a path that leads, in a general way, to an outcome the human intends, as

[77] Muḥammad b. Farāmurz (Mullā Khusrū), *Mirʾāt al-Uṣūl Sharḥ Mirqāt al-Wuṣūl*, ed. Ilyās Qablān al-Turkī (Beirut: Dār Ṣādir), 381– 401.

an instrument leads to the work performed through it. It is, in the terms of the analogy, the door, not the path that is the efficient cause. Because, in the end, the human being has the ability (maybe not *fi'lan* but *quwwatan*) to reach the goal they directed to the AI system.

The weight of an AI's *Sababiyyah* may rise with the gravity of the act it serves. For example, to operate a robotic system using an AI agent in setting a broken bone is a lighter matter than to operate one in a heart transplant; the heavier the *Taklīf*-bearing act for which the tool is employed, the more consequential its role as *sabab*. However, no rise in capability lifts the tool out of the category of *sabab*. Even the highest forms not yet realized, such as AGI or superintelligence, would remain a *sabab*: greater capability makes for a weightier occasioning cause, never an efficient one, and so can never warrant *Taklīf*.

This returns us to the central topic of the present study. Our subject is *Taklīf*, and we have already established that agentic AI can be a *Mukallaf*. Given this, the precise category into which category of tools AI is sorted into does not finally matter in Islamic law. Whether one holds it a *Sabab*, as is most accurate, or argues for some other declaratory relation, the productive responsibility, and with it the *Taklīf*, returns in every case to the human being. If even the most capable AI is a tool, the question of error cannot be evaded. A system entrusted with a medical decision may direct the wrong treatment; one relied upon in a court of law may yield the wrong judgment; one followed in finance may result in financial or commercial loss. If such a system were treated as the one responsible part, then when it erred, there would be no one to hold to account, as the system isn't *Mukallaf*. Accordingly, the legal categories through which Islamic law answers a wrong – such as blood-money (*Diyah*) and retribution (*Qiṣāṣ*) – would have nothing to attach to. Responsibility (*Mas'ūliyyah*) would be left without a subject. The logical follow-up question that Islamic law answers is: who is responsible?

Taklīf: Who Is Responsible?

Having established that AI cannot occupy the status of *Mukallaf*, two questions necessarily follow. The first is one of permissibility: is it licit to employ, in tasks that carry moral or legal accountability, a tool that is not itself capable of bearing such accountability? The second is one of attribution: given that the AI itself cannot be held responsible, who is? Responsibility must, by necessity, devolve upon a human being, since within the Sharīʿah it presupposes a legally responsible agent (*Mukallaf*) capable of intention, negligence, or transgression – qualities an AI system, as established above, cannot possess. The difficulty, however, is not whether responsibility attaches to a human agent, but rather to which human agent it attaches, particularly in cases involving multiple parties operating at different stages of the AI's design, training, and deployment. This complexity is illustrated in the case of an AI-operated surgical robot that causes the death of a patient: here, the one liable might plausibly be identified with any of at least three distinct parties – the company that manufactured and deployed the system, the engineer or research team that designed and trained its underlying model, or the practitioner who operated it and exercised (or failed to exercise) supervisory judgment during the procedure. Each of these parties stands in a different causal and epistemic relationship to the harm produced, and each may bear a different degree and kind of responsibility depending on the nature of the failure – whether it originates in defective design, inadequate training data, improper deployment, operator error, etc. The resolution of these questions, and the criteria by which responsibility is properly assigned among these parties, are addressed in what follows.

Employing *Non-Mukallaf* Systems

We would like to present the following situations that will bring the importance of such a discussion into light; the first being robotic surgery using and AI agent. According to Eric Topol and others, the use of AI has started to be introduced into AI surgeries. In the case where a trained agent is used to operate a robot for a surgery, there are two possibilities: one; it makes a mistake and the patient dies, and two: it doesn't make a mistake, and the patient still passes away due to complications. What happens? Second scenario is autonomous vehicles. Companies like Waymo are starting to use them. Two possibilities: a car malfunctions and the passenger dies, and two, the car doesn't malfunction and makes a proper decision like if a person immediately jumps in front of the car and the car choses to save both by swerving but in the process, it turned over and the passenger died. What happens? Such are real scenarios. Does this entail that this possibility of AI causing harm allude to the fact that we cannot use it?

Before proceeding to the criteria by which responsibility is properly assigned, it is worth pausing on two concrete scenarios that bring the importance of this discussion. The first concerns robotic surgery. According to Eric Topol and others, AI-assisted and AI-operated systems have increasingly been introduced into surgical practice.[78] Where a trained AI agent operates a surgical robot, two possibilities present themselves: either the system errs and the patient is harmed or dies as a direct result of that error, or the system performs without technical fault and the patient nonetheless succumbs to complications

[78] Eric Topol, "Agentic AI Comes to Medicine: Expansion of Capabilities with Two New Medical AI Models,"-Ground Truths (Substack newsletter), June 17, 2026, https://erictopol.substack.com/.

arising independently of the AI's operation. A structurally analogous dilemma arises with autonomous vehicles, now being deployed commercially by companies such as Waymo[79]: either the vehicle malfunctions and the passenger is injured or killed as a consequence, or the vehicle functions exactly as designed – for instance, executing a split-second decision to swerve in order to avoid a pedestrian who has stepped suddenly into its path – yet this very decision results in a rollover that kills the passenger. These are not merely hypothetical constructions, but scenarios already realized, or closely approximated, in contemporary deployment of such systems. They raise a question that is foundational to the remainder of this discussion: does the mere possibility that an AI system may cause harm – whether through fault or through faultless operation – entail that such systems may not be used at all, or does permissibility rest instead on a more refined set of criteria to which we now turn?

The first step is to identify if current Agentic AI models are generally accurate in performance. Empirical evidence lends considerable support to the claim that appropriately trained AI systems can perform not just adequately, but demonstrably better than their human counterparts, in domains bearing directly on human safety and well-being. In the case of autonomous vehicles, Waymo reports that, relative to an average human driver operating over the same distance in the same cities, its autonomous driving system has achieved a 94 percent reduction in crashes involving serious injury or worse, an 82 percent reduction in airbag deployments across all vehicle crashes, and an 82 percent reduction in injury-causing crashes overall. These gains extend to the protection of vulnerable road users in particular, with reported reductions of 93 percent in pedestrian crashes involving injury, 84 percent in cyclist crashes involving injury, and 84 percent in motorcycle crashes involving injury.[80] A comparable pattern emerges in medicine. In a controlled comparison reported by Eric Topol, a medical AI agentic system named MIRA achieved an overall diagnostic accuracy of 87.8 percent, compared to 78.1 percent among board-certified (BC) physicians, with the gap widening considerably for particular diagnoses – 95.2 percent versus 78.6 percent for pancreatitis, and a full 100 percent versus 88 percent for appendicitis. Although MIRA ordered a greater number of blood tests (51 percent versus 28 percent), this was offset by a substantially reduced reliance on imaging, suggesting no net increase, and possibly a net decrease, in overall resource consumption. On the therapeutic side, MIRA outperformed BC physicians in correctly ordering indicated procedures such as laparoscopic appendectomy or cholecystectomy (53.5 percent versus 38.3 percent), demonstrated superior adherence to guideline-based fluid management and analgesic protocols, and achieved an overall 35 percent higher rate of alignment with clinical guidelines. Of 468 medications MIRA ordered, 99.8 percent were correct with respect to indication and safety – including allergy status, drug interactions, and renal dosing. Notably, MIRA also triaged a greater proportion of cases for hospital admission than the physicians did, a pattern Topol interprets as reflecting the absence of economic incentive in the AI's clinical judgment.[81]

Taken together, this evidence suggests that when an agentic AI system is aligned with sufficient rigor – as Topol notes regarding MIRA, which was equipped with eleven distinct tools and a carefully curated action space – its performance can meet, and in

[79] https://waymo.com

[80] https://waymo.com/safety/

[81] Eric Topol, "Agentic AI Comes to Medicine: Expansion of Capabilities with Two New Medical AI Models."

several respects exceed, that of trained human professionals operating in the same domain. This is not a claim of mere adequacy but of demonstrated superiority along several concrete metrics of accuracy and safety, as shown above. It follows that an AI agent, when trained and aligned with sufficient rigor, may properly be characterized as a "capable" tool – one whose reliability, at minimum, need not be presumed inferior to that of the human agents it supplements or replaces.

The second question that arises is whether it is permissible to employ an entity that is not itself *Mukallaf* but has nonetheless been trained to reliably execute a given task. A useful legal parallel here is the Islamic rulings regarding the use of a trained hunting animal. The relevant legal precedent is found in the narration of ʿAdī b. Ḥātim (ra), who reports that the Prophet (saw) was

asked about hunting with trained dogs, and responded: "When you send your hunting dog after game and mention the name of Allāh, then if it catches and kills the prey, eat from it. But if it has eaten from the prey, then do not eat it, for it has only caught it for itself. And if your dog mixes with other dogs over which the name of Allāh was not mentioned, and they catch and kill the prey, then do not eat it, for you do not know which of them killed it. And if you shoot the game with your arrow and find it after a day or two, with no mark on it except the trace of your arrow, then eat it. But if it has fallen into water, then do not eat it."[82]

Several jurisprudential conditions can be extracted from this narration that bear directly on the present question.[83] First, the animal in question is not itself a moral agent – it is not *Mukallaf*, bears no independent legal responsibility, and cannot be held to account for its actions – yet its output (the caught game) is nonetheless deemed lawful for consumption, provided certain conditions are met. Second, permissibility is not unconditional but contingent upon the animal's having been properly trained (*Muʿallam*), as in obeying when sent and desisting when restrained; an untrained or improperly trained animal does not generate the same ruling. Third, the ruling is sensitive to the reliability of the causal chain between the animal's action and the outcome: where that chain is compromised – as when the dog eats from the prey, indicating it acted for its own end rather than at its owner's command, or when other, untrained dogs join the hunt such that the source of the kill becomes uncertain – the permission is revoked, notwithstanding that the animal remains, in principle, a "trained" one. What determines permissibility, then, is not the presence or absence of moral agency in the acting entity, but rather (a) whether that entity has been reliably trained for the task at hand, and (b) whether the specific instance of its action can be attributed with sufficient confidence to that training, uncontaminated by extraneous or unreliable factors.

This precedent can be also attributable to the case of AI agents. Like the hunting dog, the AI system is not *Mukallaf* and bears no independent moral or legal responsibility. And like the hunting dog, this absence of moral agency does not, by itself, render its use impermissible. What the ḥadīth establishes instead is a reliance upon a non-*Mukallaf* entity as permissible where that entity has been properly trained for the task in question and performs within the reliable bounds of that training. Applied to AI, this suggests that where an AI agent is rigorously aligned and trained – as the empirical evidence discussed above

[82] *Ṣaḥīḥ Bukhārī*, 5167.

[83] For the list of conditions regarding the hunting animal according to the four Sunni schools of law, see: Al-*Mawsūʿah al-Fiqhiyyah al-Kuwaytiyyah*, The Ministry of Awqāf and Islamic Affairs, State of Kuwait (1404-1427H0, 28:137-141).

indicates is achievable, operates within the bounds of tasks for which it has been so trained, and desists when restrained, its use is likewise permissible, notwithstanding its lack of *Taklīf*. The determinative question, in other words, is not whether the tool possesses moral agency, but whether it has been made reliably fit for the purpose to which it is put.

A third question that must be addressed is whether there is room for non-intentional error.[84] This can be of two types. The first type is when the error occurs through no fault of the AI system itself – that is, through unforeseeable circumstances rather than negligence or malfunction. The two scenarios brought before are presented. In AI-assisted surgery, the system commits no technical error, yet the patient nonetheless dies of complications unrelated to the procedure's execution. In the case of the autonomous vehicle, the system correctly identifies an obstacle and elects to swerve in order to protect the passenger, yet the vehicle overturns in the process and the passenger dies. What is the ruling in such cases? Islamic law offers a clear precedent: where a competent physician commits no act of malpractice and the patient nonetheless dies, the physician is not held liable, since the death cannot be causally or morally attributed to any deficiency in his conduct. Likewise, where a driver operates his vehicle competently and a death occurs through no fault of his own, he bears no liability.[85] By the same reasoning, an AI system that operates competently and executes the correct decision under the circumstances presented to it should not be held liable, as the *ʿIllah* (underlying legal cause) upon which liability is predicated is simply absent.

The second type of error is where genuine error does occur despite competent effort. In this case, Islamic law provides a separate category for its treatment through *Khaṭaʾ*, the legally excusable mistake, which does not give rise to liability or blame where the actor has exercised due diligence in reaching his judgment. This principle is most clearly articulated in the Ḥadīth reported by ʿAmr b. al-ʿĀṣ (ra), in which the Prophet (saw) said: "If a judge makes a ruling, striving to apply his reasoning and he is correct, he will have two rewards. If a judge makes a ruling, striving to apply his reasoning and he is mistaken, he will have one reward."[86] The Ḥadīth establishes that *Ijtihād* by a human being (who is competent in Islamic law) toward a correct outcome is rewarded regardless of whether the outcome itself proves correct. Extending this principle to AI is more complex, however, since the AI itself lacks *Taklīf*. It cannot be the bearer of reward or excuse in the way a human judge can. Rather, the relevant question becomes whether an error made by a properly functioning, diligently trained AI system should be treated, for purposes of liability, analogously to a human *Khaṭaʾ* – that is, as an excusable outcome of a sound process rather than as a chargeable fault.

For AI to be integrated into tasks of this kind at all, however, a prior and more fundamental determination must be made: whether the balance of benefit (*Maṣlaḥah*) to harm (*Mafsadah*) justifies its use in the first place. This determination proceeds along two steps. The first is identifying which tier of the *Maqāṣid al-Sharīʿah* the task in question serves – and whether it touches upon the *Ḍarūriyyāt* (necessities), the *Ḥājiyyāt* (needs), or the *Taḥsīniyyāt* (superfluities) – since the strength of the case for permissibility is according to the weight of the interest being served. The second asks how certainly the benefit in

---

[84] As harm that is done intentionally is consequential without a doubt.

[85] There is a difference between *Taklīf* and *Ḍamān* (liability). Liability is *Waḍʿī*, as in it can attach to something that is not *Mukallaf* (as is shown with the hunting animal). *Taklīf* refers to liability that is consequence both in this world as well as in the hereafter.

[86] *Ṣaḥīḥ Bukhārī*, 7352.

question is actually attainable: whether it is *Mutayaqqan* (certain), *Mutarajjiḥah* (preponderant), *Ghālib al-Ẓann* (of strong likelihood), merely *Mumkin* (possible), or *Mawhūm* (speculative) – since a *Maṣlaḥah* that is only speculative cannot outweigh a *Mafsadah* that is concrete. It then follows that one must do the same with the *Mafsadah* and weigh the outcomes in order to develop a ruling on the permissibility of its use.

Applying this metric to the cases at hand as an example, the evidence clarified above indicates that the benefit is not merely speculative but approaches certainty (*Mutayaqqan*). AI's demonstrated accuracy in reducing surgical malpractice, in extending competent medical care to underserved populations – including populations in conflict or crisis zones such as Gaza and Sudan who otherwise lack access to physicians – and in reducing traffic accidents through autonomous vehicles, together serve one of the five universally recognized objectives of the Sharīʿah, *Ḥifẓ al-Nafs* (the preservation of life), and does so at the level of a *Maṣlaḥah ʻĀmmah* (a public benefit). Having established the weight of the *Maṣlaḥah*, the relevant legal maxim is applied: *Darʼ al-Mafāsid Awlā min Jalb al-Maṣāliḥ* – "preventing harms takes precedence over procuring benefits." Applying this maxim then requires the identification of how certain and how severe the corresponding *Mafsadah* actually is, and how it compares to the rate and severity of harm already produced by human practitioners operating in the same domain without AI assistance. This requires setting the error and harm rates of medical AI agents against equivalent rates for human physicians – including not only diagnostic error rates, but also documented rates of physician misconduct, negligence, and ethical violation – drawing on credible research in each case.[87] When this comparison is carried out, the balance of evidence points toward the conclusion that, for surgical intervention and for autonomous driving specifically, the *Maṣāliḥ* substantially outweigh the *Mafāsid*.

The same conclusion does not hold uniformly across all applications of AI, however. In domains such as military targeting or mass surveillance, the potential for unethical use is sufficiently high, and sufficiently severe in its consequences, that the *Mafāsid* there run parallel to, or outweigh, the *Maṣāliḥ*. This should also not be taken to imply that any domain of AI application is entirely free of *Mafsadah*: an AI agent tasked with preparing food may simply burn it, resulting in a loss against *Ḥifẓ al-Māl* (the preservation of wealth), just as any human error would. The relevant threshold is not the mere existence of a possible *Mafsadah*

[87] The overall diagnostic error rate across all diseases is estimated at 11.1 percent – roughly one in nine diagnoses rendered incorrect – with some 12 million Americans misdiagnosed annually, and diagnostic error implicated in nearly a quarter of cases resulting in death or ICU transfer. See: David E. Newman-Toker, Najlla Nassery, Adam C. Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D. Clemens, Zheyu Wang, Yuxin Zhu, Ali S. Saber Tehrani, Mehdi Fanai, Ahmed Hassoon, and Dana Siegal, "Burden of Serious Harms from Diagnostic Error in the USA," *BMJ Quality & Safety* 33, no. 2 (2024): 109–120, https://doi.org/10.1136/bmjqs-2021-014130; Andrew D. Auerbach, Tiffany M. Lee, Colin C. Hubbard, Sumant R. Ranji, Katie Raffel, Gilmer Valdes, John Boscardin, Anuj K. Dalal, Alyssa Harris, Ellen Flynn, and Jeffrey L. Schnipper, for the UPSIDE Research Group, "Diagnostic Errors in Hospitalized Adults Who Died or Were Transferred to Intensive Care," JAMA Internal Medicine 184, no. 2 (2024): 164–173, https://doi.org/10.1001/jamainternmed.2023.7347; Melinda J. Helbock, "America's Hidden Epidemic: The Deadliest Misdiagnoses Forecast 2026," Law Firm of Melinda J. Helbock, A.P.C., last modified April 2026, https://www.helbocklaw.com/americas-hidden-epidemic-the-deadliest-misdiagnoses-forecast-2026/.

– since some possibility of harm attends virtually every action, whether human, animal, or AI – but rather whether there exists a genuine *Iḥtimāl* (likelihood) that the direct and foreseeable result of the action will be a *Mafsadah* of real consequence. Where such a likelihood exists, as arguably in the cases of surveillance, military application, or large-scale displacement of human employment, the matter falls under the principle of *Sadd al-Dharā'i*ʿ (blocking the means to harm), and the application should accordingly be avoided or heavily restricted, notwithstanding the technology's demonstrated capability elsewhere.

Lastly comes the question of liability. Even where the *Maṣlaḥah-Mafsadah* metric favors the use of AI, and even where the risk of harm has been shown to be genuinely lower than the corresponding human baseline, a further question remains outstanding: in the event that a mistake attributable to the AI does occur, who bears *Taklīf*, and to whom does responsibility attach? This question is addressed in the section that follows. It suffices for now to note that, regardless of how low the residual risk associated with AI may be shown to be, some party must nonetheless be identifiable as liable, in proportion to the type of failure the AI in question produced. It is precisely for this reason that an AI agent cannot be permitted to operate as though it possessed full, unsupervised agency: to do so would render the assignment of liability – in the cases where liability is genuinely warranted – practically impossible.

The Direct Agent and the Indirect Causer

A foundational *Uṣūlī* principle invokes directly on the question of where responsibility properly attaches: *ghayr al-mukallaf lā yuḍāf ilayhi* – "no legal ruling is ascribed to one who bears no *Taklīf*." The force of this principle, and its limits, are best illustrated as follows: consider the case of a person who pays an assassin to end another individual's life: should the assassination occur, it is the assassin – not merely the one who commissioned it – who is held liable for the killing. This is because the assassin, apart from having been hired and instructed, retained complete *Irādah* (independent will) and *Taklīf* in deciding whether to carry out the act; the assassin's own agency intervenes between the instruction given and the harm caused, and it is this intervening agency that anchors liability to the assassin himself. This precedent, however, cannot be transferred onto the case of AI, because AI does not possess the *Irādah* or *Taklīf* that grounds the assassin's independent liability. The Islamic rulings that connect here more are that of the hunting animal considered earlier: where a trained hunting dog kills another person's sheep, it is the dog's owner – not the dog – who is held liable, precisely because the dog, lacking *Taklīf*, cannot itself bear responsibility for having acted inappropriately. Liability, in such a case, reverts entirely to the human party who deployed the animal.

From these two contrasting precedents, a governing principle emerges. Where the acting entity is merely a tool (*sabab*) employed to carry out a task, and lacks independent *Taklīf* of its own, liability for harm arising from that task transfers upon the human party responsible for its deployment – not upon the tool itself, regardless of how sophisticated or capable that tool may be. If an AI agent is used for a task it is capable of executing improperly – for instance, if it is deployed in a manner that results in the mistreatment of a patient – the human party who deployed it bears liability for that outcome. Likewise, where the AI was not properly aligned, was not supplied with correct or sufficient data, through the negligence of its trainer or operator, liability falls upon that party – not because the AI acted wrongly in any morally chargeable sense, but because a human party failed in the duty of care owed in bringing the tool to a reliable standard before its deployment. In every

conceivable scenario, then, the AI itself can never be responsible, since it is not, and cannot be, *Mukallaf*. What remains as a concern is not whether liability transfers to a human party (this much has now been established). Rather, given that there are more than one set of human parties potentially involved at different stages of the AI's design, training, and deployment, precisely which among them bears liability.

The relevant *Qāʿidah Fiqhiyyah* is: *idhā ijtamaʿa al-mubāshir wa-l-mutasabbib, uḍīfa al-ḥukm ilā al-mubāshir* – when the direct agent (*al-mubāshir*) and the indirect causer (*al-mutasabbib*) coincide, the ruling is ascribed to the direct agent.[88] The Muslim jurist Al-Ḥamawī defines the *mubāshir* as the one from whose own act the destruction (*talaf*) results, with no intervening act interfering between his act and the harm. The direct agent, in other words, is the last free hand to touch the chain (*al-Waṣf al-Akhīr*) before the harm occurs. It is precisely this "no intervening act" requirement, however, that complicates the maxim's application to agentic AI: where such a system executes an extended, adaptive chain of its own action between the operator's initiation and the eventual harm, the operator's act is no longer the last thing to touch the chain before harm results – the AI's own processing is. Whether that processing counts as an intervening act in the sense the maxim requires, given that it issues from an entity that is not *Mukallaf*, is the question the remainder of this section must resolve.

The classical examples in the books of *Uṣūl al-Fiqh* and *al-Qawāʿid al-Fiqhiyyah* make this rule concrete. For example, one who digs a well incurs no legal liability for a death caused by another who pushes a person into it – even if the well was itself trespassing property. This is because digging on its own doesn't necessitate destruction; the harm arises only when the act of pushing supervenes, and the pushing, as the final qualifying act (*al-waṣf al-akhīr*), is what the ruling follows. Another example cited by jurists involves a ship fully laden with cargo. If an individual casts even one pound of cargo that causes the ship to sink, the last individual to put the cargo is legally liable. In yet another legal case, one who merely directs a thief to another's property bears no liability for the theft the thief commits. In each case the occasioning cause (*sabab*) may lay the foundations for a certain consequence, but the ruling attaches to the last direct agent whose free act actually brought the consequence about.

Applied to AI, however, the connection is not as clean as it may first appear. The designer who builds the AI system and the engineer who trains it clearly stand to the harm as *Mutasabbib*, or the indirect causers who furnish the occasion, at a remove from the eventual outcome. The operator, by contrast, cannot simply be assigned the position of *Mubāshir* by default. Al-Ḥamawī's definition requires that the *Mubāshir*'s own act produce the harm with nothing intervening; but between the operator's act of initiating an agentic system and the harm that eventually results lies the system's own extended processing, which acts as a genuine intervening step, even if not one performed by a *Mukallaf*. In some scenarios, this means there may be no human *Mubāshir* at all: the last "hand" to touch the chain before harm occurs is the AI's own operation, not the operator's. This is the difficult case the maxim was never built to address, since the maxim presupposes that a human intervening act severs the causal chain running back to the *Mutasabbib*, and it is precisely this presupposition that agentic AI unsettles.

[88] Abū Bakr b. al-Mullā al-Aḥsāʾī, *Zawāhir al-Qalāʾid ʿalā Muhimmāt al-Qawāʿid*, 2nd ed. (Beirut: Dār al-Kutub al-ʿIlmiyyah, 2013), 140–142. The definition of *al-mubāshir* and the illustrations that follow are drawn from the same passage. Also, see Majallah al-Aḥkām al-ʿAdliyyah (Beirut: Dar Ibn Hazm), 2004.

The more defensible route, accordingly, is to treat the operator not as an automatic *Mubāshir* but as a *Mutasabbib* in his own right – one among several (alongside the designer and trainer) – whose liability turns not on temporal proximity to the harm but on *Ta'addī* (transgression of one's proper bounds), along the lines already established in the well-digger case above. On this reading, the operator of an AI-run surgical robot is liable not because he was the one who initiated the procedure, but specifically where his conduct constituted *Ta'addī* or *Tafrīṭ* – deploying the system outside its validated use, ignoring known limitations, or failing to supervise where supervision was warranted. Where his conduct falls short of that standard, liability may properly rest instead with the designer or trainer whose own *Ta'addī* or *Tafrīṭ* was the operative cause.

Within this fault-based framework, jurists have presented a few exceptions to the maxim that are related to the present study. One such exception that we found that relates the present study is the case of deception (*Ghurūr*). In such cases, liability follows the deceiver rather than the direct agent. The case brought in the works of *Uṣūl al-Fiqh* concerns a woman's guardian (*Walī*) or agent (*Wakīl*) who tells a man, "Marry her, for she is a free woman"; the man marries her and she bears him a child, whereupon it emerges that she was in fact another's slave. The deceived party (*Maghrūr*) is granted recourse for the value of the child against the one who deceived him. The principle drawn from it is that the one who perpetrated the *Ghurūr* bears the consequence, for the deceived party's own conduct does not rise to the level of *Ta'addī* or *Tafrīṭ* – he acted reasonably upon the false picture the deceiver had set before him, and reasonable reliance is not negligence.

This transfers cleanly to AI. In cases where an Agentic AI system is used to operate a medical robot in a life-saving procedure and the patient dies, or is trained to present legal cases in a court of law and render an erroneous judgement, or is trained to make financial decisions and incurs financial or commercial losses, the practitioner who operated it – or who most directly initiated its function – bears the responsibility where his deployment of the system itself constituted *Ta'addī* or *Tafrīṭ*. But where the company that supplied the system knowingly advertised it falsely, such as marketing it for a practice it was not fit to perform or concealed certain major problems from its customers – the company has committed *Ghurūr*, and the liability shifts to it. The operator who acted in good faith upon a misrepresented tool stands in the position of the *Maghrūr*; it is the deceiver who answers for the harm.

Another exception made is that of transgression or wrongful exceeding of one's bounds (*Ta'addī*) considered on its own terms, which in fact supplies the general rule of which the operator's liability above is simply an instance. The case brought in the works of *Uṣūl al-Fiqh* concerns one who directs a thief to a property that they themselves were in charge of protecting. In the ordinary case where they merely direct a thief to another's property, they bear no liability for the theft the thief then commits, since they are no more than a *Mutasabbib*. But where they were responsible for that property's protection, they have now violated their responsibility and thereby committed *Ta'addī*; and although they remain, as to the theft itself, a *Mutasabbib*, this secondary act of transgression is what renders them legally liable.

This transfers cleanly to the case of AI. Where an agentic AI system is employed for unlawful surveillance, a company that supplied the underlying data is not held liable if its terms of service had expressly disclaimed any undertaking to protect its users' stored information. Thus, the user, in agreeing, bore the risk, and the company occasioned the harm only as a *Mutasabbib* without *Ta'addī*. But where the company had promised to safeguard its users' information and instead handed that data to the designer who built the

surveillance system, it has breached its responsibility of protection and thereby committed *Ta'addī*, and the *ḍamān* therefore attaches to it.

When the User is Liable

This fault-based framework – in which liability among mutasabbib parties turns on *Ta'addī* and *Tafrīṭ* rather than on which party happened to act last – carries forward into a final scenario, arguably the one of greatest consequence for the future trajectory of Agentic AI. The tools currently available to AI agents are still relatively constrained, such as web search, scheduling, and form completion. The agentic trajectory points toward AI systems wielding increasingly powerful and consequential tools, undertaking tasks far beyond narrow commands like "perform this surgery" or "drive this vehicle from point A to point B." It will not be long before such agents are making purchases, planning trips, evaluating medical options, and acting not merely on explicit instruction but on their own inference of what the user needs – extending, in effect, into "full autonomy" on the user's behalf. What is new, and rapidly approaching, is its migration into ordinary personal life. As this occurs, the operative distance between the AI's creator and its end user grows: the system's behavior becomes progressively less a direct extension of its designer's original specification and progressively more a function of how, and for what, the individual user chooses to deploy it.

This trajectory finds an instructive parallel in the ḥadīth reported by Abū Hurayrah (ra), in which the Prophet (saw) said: "There is no compensation for one killed or wounded by an animal or by falling in a well or because of working in a mine."[89] The juristic elaboration on this principle, as recorded by the ḥadīth scholar al-Bukhārī, contains the opinions of the scholars of the *Salaf*: Ibn Sīrīn held that no compensation was owed for an animal's kick, unless someone had injured the animal while directing its reins, in which case liability shifted to the rider. Ḥammād similarly held that a kick by the animal incurred no liability, except where someone had incited the animal, in which case liability fell upon the instigator. Shurayḥ ruled that where retaliation had occurred – that is, where a person had struck the animal first and was then kicked in return – no compensation was due. Al-Ḥakam addressed the case of a laborer driving a donkey carrying a woman: if she fell, no liability attached to the laborer. And al-Sha'bī distinguished between a driver who exhausted the animal through his own conduct, thereby incurring liability for resulting harm, and one who drove it gently and in the customary manner, who did not.

What emerges from this body of juristic elaboration is not a blanket rule of non-liability for harm caused by an intermediary agent, but a more precise principle: liability tracks the human party's conduct in relation to that intermediary – its provocation, its mistreatment, its being placed under conditions likely to produce harm – rather than attaching automatically to the intermediary itself, and rather than being waived automatically simply because the immediate cause of harm was non-human. AI stands in an analogous, if intensified, position. It operates far closer to a human being in capacity than any animal does, and yet, in relation to *Taklīf*, it remains what the animal is: a *sabab* through which harm may occur, rather than a bearer of independent moral responsibility. Crucially, this holds even where the AI's behavior diverges from the alignment it was originally given, in ways that cannot be traced back to any specific human input or negligence in the moment of the error itself. The human party who deployed it remains

[89] *Ṣaḥīḥ Bukhārī*, 6912.

liable regardless, precisely because that party knowingly introduced an instrument whose capacity for error was foreseeable in principle, even if its particular manifestation was not foreseeable in the specific instance.

Consider a doctor who employs an autonomous AI agent to manage his hospital's administrative affairs without direct human oversight, and the agent errs in registering a patient's appointment in a manner that results in the patient's death. The fault, on this analysis, remains the doctor's – not because he committed any error in the moment, but because he was the one who employed the system with knowledge of its inherent possibility of error, much as al-Sha'bī's driver bears liability for exhausting his animal even without directly causing the resulting harm by his own hand.

Two exceptions to this rule of user liability follow directly from the ḥadīth's own logic. The first mirrors the case of one who falls into a well or is wounded by a stray animal through his own action: where the harm results from the fault of the affected party himself – as where a patient fails to schedule his appointment, or fails to arrive for treatment as instructed – liability does not attach to the one who deployed the AI, since the causal chain terminates in the victim's own conduct rather than in the deployer's decision to employ the system. The second concerns cases where an AI agent is deployed upon another individual with that individual's informed consent – a state of *Riḍā Bayn al-Jānibayn*, mutual agreement between the two parties. In such cases, the party subjected to the AI's operation has voluntarily relinquished his right to hold the deployer accountable for the risks he knowingly accepted.

From this follows a governing principle of use: the greater the *Taklīf*-bearing weight of an action, the more cautious one must be upon AI to perform it. Employing AI to draft a presentation for school is a light matter; but to employ AI in operating a robot during a surgical procedure is a grave matter, for an error there may cost a life that cannot be restored. The first invites free use; the second demands the closest scrutiny, and at its upper limit may require that the human carry out the act himself rather than delegate it to a tool whose processing he cannot fully inspect. To calibrate reliance to the gravity of the task is, in the end, an obligation upon the *Mukallaf*, who answers for whatever they set in motion through the *Sabab*. The precaution must scale with the gravity of the task.

Conclusion: Accounting for Macro and *Sadd al-Dharā'i*ʿ

As a closing reflection, a true legal analysis of AI must also move from the individual act to the macro – namely, the consequences a ruling can set in motion once it is generalized. A *Fatwā* is never a merely private permission; it licenses a practice across a whole community, and at that scale effects emerge that the isolated case never had. A ruling must therefore be weighed not only by its immediate content, but by what it opens up. This is the proper domain of *Sadd al-Dharā'i*ʿ – the blocking of the means: the principle that an act otherwise permissible may be prohibited or at least narrowed in its permissibility when it reliably leads to harm. Applied to AI, it directs attention past the immediate question of "may this tool be used?" to the secondary and tertiary effects that widespread, sanctioned use would unleash.

The Fear of Transhumanism and the Erosion of Human Capability

A friend once related how he explained artificial intelligence to his father, a man with no background in the subject. He asked him, "When you call someone, do you know exactly how your voice reaches the other person?" His father replied, "No." "Do you still use the phone?" "Yes." "Well," he said, "that is the same with AI – except it will be with our thoughts."

The exchange captures the particular anxiety this final section addresses: not merely that AI might err, or that its use raises questions of liability, but that its habitual use may begin to reshape our own capacity to think and act independently. This is, at its root, a fear of transhumanism. The human being possesses a specific constitution whose faculties carry an intended purpose, and that intervention which degrades or displaces those faculties runs contrary to the wisdom of their creation. The intellectual and spiritual faculties that are, no less than the body, part of the *Fiṭrah* Allah has given the human being. If permanently and artificially altering the body is problematic because it interferes with a form Allah deemed already complete (as usually discussed in transhumanism, i.e. genetic engineering and alteration), habitually delegating the faculty of thought to an external system raises a similar concern; not the alteration of a physical form, but the deterioration of a spiritual and intellectual one through over-outsourcing.

This concern is no longer speculative. A 2025 study by Michael Gerlich, published in Societies and drawing on a mixed-methods survey of 666 participants across a range of ages and educational backgrounds, found a measurable negative relationship between frequent AI tool use and critical thinking ability, mediated specifically by what researchers term cognitive offloading – the practice of delegating mental effort to an external tool rather than exercising it oneself.[90] The study further found that younger users, who exhibit the heaviest reliance on AI tools, showed correspondingly lower critical thinking performance, while higher education served as a partial buffer against this decline. What the friend's analogy captures intuitively, in other words, the empirical literature now confirms directly: sustained delegation of cognitive effort to AI does not leave the underlying faculty

[90] Michael Gerlich, "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," *Societies* 15, no. 1 (2025): article 6, https://doi.org/10.3390/soc15010006

untouched but measurably diminishes it.

A further and more pressing concern is that the loss of logical autonomy does not remain contained in the intellectual domain, but threatens to extend into a loss of spirituality as well. The scholar al-Ghazālī's discussion of gratefulness can be applied here.[91] Gratefulness, as al-Ghazālī insists, is not a matter of verbal acknowledgment but of using a given faculty according to the wisdom for which it was created; he illustrates this through the example of wealth, whose purpose is to facilitate the meeting of human need, such that hoarding it constitutes a failure of gratitude despite its outward preservation. The human intellect (*ʿAql*) stands in a closely similar position. Its purpose is to struggle against base inclinations and to arrive, through sustained reflection (*Tafakkur*) and effort, at the knowledge of Allāh and sound moral judgment.[92] An intellect that is preserved in the sense of never being damaged, yet is never actually exercised – because every act of reflection, evaluation, and judgment has been handed over to an external system – is, on this logic, no more being used according to the wisdom of its creation than hoarded wealth is. The faculty remains intact in name while its purpose goes unfulfilled. And if that faculty is progressively left unexercised through habitual delegation to AI, the concern is not merely philosophical but practical: a mind unaccustomed to independent reflection may become less equipped, over time, to exercise precisely the kind of sound, spiritually anchored judgment that *Taklīf* presupposes.

Conclusion

There remain, several additional concerns that could sustain an argument for a stricter application of *Sadd al-Dharāʿi*ʿ to AI than has been developed here – among them, the erosion of privacy,[93] the unethical use of AI in surveillance and manipulation, sexual exploitation, the large-scale displacement of human labor, and the deepening of existing socio-economic imbalances as access to and control over such systems concentrates among a narrower set of actors. A full treatment of each lie beyond the scope of this paper, but their cumulative weight is worth noting that as AI grows more advanced, it is not only its benefits that scale accordingly, but its potential harms as well.

---

[91] Abū Ḥāmid al-Ghazālī, *Iḥyāʾ ʿulūm al-dīn* (Beirut: Dār al-Fikr, 2003), 4:79

[92] Abū Ḥāmid al-Ghazālī, *Iḥyāʾ ʿulūm al-dīn,* 4:190.

[93] China's mass-surveillance system in Xinjiang treats the ordinary markers of Muslim religiosity – such as the keeping of a beard, religious attire, the presence of the Qurʾān on a person's phone – as indicators of "extremism." The system feeds these criteria into an algorithmic system that flags Turkic Muslims for interrogation and detention. The firm Clearview AI assembled a database of more than thirty billion facial images, scraped indiscriminately from social media and the open web without the knowledge or consent of those depicted, and converted each face into a searchable biometric "faceprint" sold on to law-enforcement clients; European regulators have repeatedly found the practice unlawful, and the United Kingdom's Information Commissioners' Office (ICO) fined the company some €7.5 million for what it called an "illegal database." See: Human Rights Watch, *China's Algorithms of Repression: Reverse Engineering a Xinjiang Police Mass Surveillance App* (New York: Human Rights Watch, 2019). The report documents how the Integrated Joint Operations Platform flags lawful religious expression and appearance as indicators of "extremism"; Chris Vallance, *Face search company Clearview AI overturns UK privacy fine*, BBC, 2023. URL: https://www.bbc.com/news/technology-67133157

There is, without question, genuine progress to be found in these technologies, as the evidence brought throughout this paper attests to that. But progress of this kind is rarely unambiguous, and it is worth asking whether as many bridges are being burned in the process as are being built. Not all progress is *good* progress simply by virtue of being progress; advancement in capability is not identical to advancement in human welfare. At some point, the question must be asked: is this particular application of AI genuinely solving a problem that exists, or is it, in the process, quietly creating new ones in its place?

None of these warrants abandoning AI. The relevant distinction – after discussing AI as an accurate tool and then its use and measure of responsibility – is between AI used as an aid to human reflection and AI used as its wholesale replacement. Just as therapeutic medical intervention restores or supports the body's given constitution without supplanting it, AI can be used to inform, expedite, and support human judgment without displacing the underlying exercise of *Tafakkur*, *Ijtihād*, and independent moral reasoning that the *Fiṭrah* both enables and requires. The practical safeguard, then, is proportion. Consulting one's own reflection before or alongside consulting the machine, treating its output as an input to one's judgment rather than a substitute for it, and remaining alert – as the empirical evidence now suggests one must – to the difference between a tool that extends human capability and one that quietly, gradually, comes to "replace" it.

It is safe to say that with AI's rapidly evolving nature, a general ruling on AI cannot be made, and rather it is that every situation must be handled on its own. It is for this sheer number of entangled threads within artificial intelligence, each dragging its own consequences behind it – that no single, concrete *Fatwā* can responsibly be issued upon AI as such. What we discuss here is a possible framework to measure each situation regarding its permissibility of use using the Qur'ān, the Sunnah, legal theory (*Uṣūl al-Fiqh*), and the objectives of Islamic law (*al-Maqāṣid al-Sharīʿah*). Additionally, we provide the theological groundwork; that AI is an instrument – a *Sabab*; it cannot be *Mukallaf*; and responsibility returns in every case to the human being. But the further question of what exactly may be permitted, and under which conditions, cannot be settled in the abstract, for the macro consequences shift with every use. What can be offered apart from the legal framework above is, that where the means tend reliably toward harm, caution must be expressed.

Bibliography

Advenae, Aurum. "When LLMs Choose Blackmail and Murder over Shutdown: Anthropic's Agentic Misalignment." *AI Connect Network*. 2025. https://aiconnectnetwork.com/anthropic-report-on-llm-blackmail/.

al-Aḥsāʿī, Abū Bakr ibn al-Mullā. *Zawāhir al-Qalāʾid ʿalā Muhimmāt al-Qawāʿid*. 2nd ed. Beirut: Dār al-Kutub al-ʿIlmiyyah, 2013.

al-Ālūsī, Shihāb al-Dīn. *Rūḥ al-Maʿānī fī Tafsīr al-Qurʾān al-ʿAẓīm*. Beirut: Dār Iḥyāʾ al-Turāth al-ʿArabī, 1994.

Anthropic. "Agentic Misalignment: How LLMs Could Be Insider Threats." June 20, 2025. https://www.anthropic.com/research/agentic-misalignment.

Anthropic. "Claude's Constitution." https://www.anthropic.com/constitution.

Arozullah, Ahsan M., and Mohammed Amin Kholwadia. "Wilāyah (Authority and Governance) and Its Implications for Islamic Bioethics: A Sunnī Māturīdī Perspective." *Theoretical Medicine and Bioethics* 34, no. 2 (2013): 95–104.

Auerbach, Andrew D., Tiffany M. Lee, Colin C. Hubbard, Sumant R. Ranji, Katie Raffel, Gilmer Valdes, John Boscardin, Anuj K. Dalal, Alyssa Harris, Ellen Flynn, and Jeffrey L. Schnipper, for the UPSIDE Research Group. "Diagnostic Errors in Hospitalized Adults Who Died or Were Transferred to Intensive Care." *JAMA Internal Medicine* 184, no. 2 (2024): 164–173.

al-Bāqillānī, Abū Bakr Muḥammad ibn al-Ṭayyib. *al-Taqrīb wa-al-Irshād (al-Ṣaghīr)*. Edited by ʿAbd al-Ḥamīd ibn ʿAlī Abū Zunayd. 2nd ed. Beirut: Muʾassasat al-Risālah, 1998.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. "On the Opportunities and Risks of Foundation Models." arXiv preprint arXiv:2108.07258 (2021). https://arxiv.org/abs/2108.07258.

al-Bukhārī, Muḥammad ibn Ismāʿīl. *Ṣaḥīḥ al-Bukhārī*.

Burtell, Matthew, et al. "The Surprising Power of Next-Word Prediction: Large Language Models Explained, Part 1." Center for Security and Emerging Technology (CSET). https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/.

Choi, Rene Y., Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang, and J. Peter Campbell. "Introduction to Machine Learning, Neural Networks, and Deep Learning." *Translational Vision Science & Technology* 9, no. 2 (2020): article 14.

al-Dihlawī, Shāh Walīullāh. *Ḥujjat Allāh al-Bālighah*. Beirut: Dār Ibn Kathīr, 2020.

Gerlich, Michael. "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking." *Societies* 15, no. 1 (2025): article 6.

al-Ghazālī, Abū Ḥāmid. *Iḥyāʾ ʿUlūm al-Dīn*. Beirut: Dār al-Fikr, 2003.

Goertzel, Ben. "Artificial General Intelligence: Concept, State of the Art, and Future Prospects." *Journal of Artificial General Intelligence* 5, no. 1 (2014): 1–48.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press, 1985.

Helbock, Melinda J. "America's Hidden Epidemic: The Deadliest Misdiagnoses Forecast 2026." Law Firm of Melinda J. Helbock, A.P.C. Last modified April 2026. https://www.helbocklaw.com/americas-hidden-epidemic-the-deadliest-misdiagnoses-forecast-2026.

Human Rights Watch. *China's Algorithms of Repression: Reverse Engineering a Xinjiang Police Mass Surveillance App*. New York: Human Rights Watch, 2019.

Ibn Kathīr, Abū al-Fidāʾ Ismāʿīl. *al-Sīrah al-Nabawiyyah*. Extracted from *al-Bidāyah wa-al-*

Nihāyah. Edited by Muṣṭafā ʿAbd al-Wāḥid. Beirut: Dār al-Maʿrifah li-al-Ṭibāʿah wa-al-Nashr wa-al-Tawzīʿ, 1976.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013.

LangChain. "State of AI Agents." June 12, 2026. https://www.langchain.com/stateofaiagents.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature* 521 (2015): 436– 444.

MacDiarmid, Monte, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. *Natural Emergent Misalignment from Reward Hacking in Production RL*. Anthropic and Redwood Research, November 23, 2025. https://arxiv.org/abs/2511.18397.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine* 27, no. 4 (2006): 12–14.

Mitchell, Tom M. *Machine Learning*. New York: McGraw-Hill, 1997.

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. "Levels of AGI for Operationalizing Progress on the Path to AGI." arXiv preprint arXiv:2311.02462 (2023). Published in *Proceedings of the 41st International Conference on Machine Learning* (2024).

Mullā Khusrū, Muḥammad ibn Farāmurz. *Mirʾāt al-Uṣūl Sharḥ Mirqāt al-Wuṣūl*. Edited by Ilyās Qablān al-Turkī. Beirut: Dār Ṣādir, n.d.

Muslim ibn al-Ḥajjāj. *Ṣaḥīḥ Muslim*.

al-Namlah, ʿAbd al-Karīm ibn ʿAlī ibn Muḥammad. *al-Muhadhdhab fī ʿIlm Uṣūl al-Fiqh al-Muqāran*. Riyadh: Maktabat al-Rushd, 1999.

Newman-Toker, David E., Najlla Nassery, Adam C. Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D. Clemens, Zheyu Wang, Yuxin Zhu, Ali S. Saber Tehrani, Mehdi Fanai,
Ahmed Hassoon, and Dana Siegal. "Burden of Serious Harms from Diagnostic Error in the USA." *BMJ Quality & Safety* 33, no. 2 (2024): 109–120.

Ng, Andrew. "Agentic Design Patterns, Part 1: Four AI Agent Strategies That Improve GPT-4 and GPT-3.5 Performance." *DeepLearning.AI: The Batch*, 2024. https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/.

Nisa, Ume, Muhammad Shirazi, Mohamed Ali Saip, and Muhammad Syafiq Mohd Pozi. "Agentic AI: The Age of Reasoning—A Review." *Journal of Automation and Intelligence* 5 (2026): 69–89.

The Qurʾān.

al-Raysūnī, Aḥmad. *Naẓariyyat al-Maqāṣid ʿinda al-Imām al-Shāṭibī*. 2nd ed. Riyadh: al-Dār al-ʿĀlamiyyah lil-Kitāb al-Islāmī, 1992.

al-Rāzī, Fakhr al-Dīn. *Mafātīḥ al-Ghayb (al-Tafsīr al-Kabīr)*. Beirut: Dār al-Fikr, 1981.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson, 2021.

al-Ṣanʿānī, Muḥammad ibn Ismāʿīl al-Amīr. *Irshād al-Nuqqād ilā Taysīr al-Ijtihād*. Kuwait: al-Dār al-Salafiyyah, 1985.

Shihab, M. Quraish. *Tafsīr al-Miṣbāḥ*. Jakarta: Lentera Hati, 2009.

Silfverskiöld, Ida. "Agentic AI: Comparing New Open-Source Frameworks." April 15, 2025. https://www.ilsilfverskiold.com/articles/agentic-ai-comparing-new-open-source-frameworks.

Stryker, Cole, and Mark Scapicchio. "What Is Generative AI?" IBM. https://www.ibm.com/think/topics/generative-ai.

Sutton, Richard. "The Bitter Lesson." *Incomplete Ideas* (blog). March 13, 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

al-Tirmidhī, Muḥammad ibn ʿĪsā. *Jāmiʿ al-Tirmidhī*.

Topol, Eric. "Agentic AI Comes to Medicine: Expansion of Capabilities with Two New Medical AI Models." *Ground Truths* (Substack newsletter), June 17, 2026. https://erictopol.substack.com/p/agentic-ai-comes-to-medicine.

Vallance, Chris. "Face Search Company Clearview AI Overturns UK Privacy Fine." *BBC News*, October 18, 2023. https://www.bbc.com/news/technology-67133157.

Waymo. https://waymo.com.

Waymo. "Safety." https://waymo.com/safety/.

Wizārat al-Awqāf wa-al-Shuʾūn al-Islāmiyyah (Kuwait). *al-Mawsūʿah al-Fiqhiyyah al-Kuwaytiyyah*. Kuwait: Wizārat al-Awqāf wa-al-Shuʾūn al-Islāmiyyah, 2006.

Wooldridge, Michael, and Nicholas R. Jennings. "Intelligent Agents: Theory and Practice." *Knowledge Engineering Review* 10, no. 2 (1995): 115–152.

Xu, Bowen. "What Is Meant by AGI? On the Definition of Artificial General Intelligence." arXiv preprint arXiv:2404.10731 (2024).

Zewe, Adam. "Q&A: What Is Agentic AI Today, and What Do We Want It to Be?" *MIT News*, June 30, 2026. https://news.mit.edu/2026/agentic-ai-and-what-do-we-want-it-be-0630.