Discover Artificial Intelligence

Discover

Perspective

Islamic virtue-based ethics for artificial intelligence

Amana Raquib[1] · **Bilal Channa**[1] · **Talat Zubair**[2] · **Junaid Qadir**[3]

Received: 14 March 2022 / Accepted: 7 June 2022
Published online: 20 June 2022

Abstract
The twenty-first century technological advances driven by exponential rise of artificial intelligence (AI) technology have ushered in a new era that offers many of us hitherto unimagined luxuries and facilities. However, under the guise of this progressive discourse, particularly in the backdrop of current neo-liberal late-capitalist postmodern world, AI development also has prompted an increasingly uncertain ethical tomorrow. This paper aims to probe the question of ethics by exploring the true ramifications of AI and interrogating its various ethical dimensions. It questions the essential goodness that is attributed to unstinted AI development before elucidating the ethical repercussions of AI advancements and the aptness of the current market logics and business models that govern the tech-industry. The paper next positions a holistic Islamic virtue-based AI ethics framework grounded in the context of Islamic objectives (*maqāṣid*) as an alternative ethical system for AI governance. We argue that this distinctive Islamic virtue-based ethical approach, which can be used to explore AI-related ethical problems more holistically due to its ontological base and rich tradition while keeping in check undue influence from the current socio-politico-economic climate, can be a valuable addition to the global discourse on AI ethics.

Keywords AI ethics · Islamic ethics

1 Introduction

Recent advances in machine learning (ML) and artificial intelligence (AI), driven by the ability to process large amounts of data, to learn patterns and make predictions has turbo charged the potency of traditional technology. In contrast to an overly optimistic view of technology driven by a simplistic progressivist attitude towards science and technology sees technological progress as necessary, even inevitable, for the next leap of the evolution of the *Homo sapiens*, there are now growing concerns over the harms that can be wrought by these technologies (referred to as "weapons of math destruction" by Cathy [35]. As documented by Vinuesa et al. [52] and Latif et al. [25], AI technology based on data processing is not neutral or without harmful concomitants and can both promote or inhibit human development [25, 38].

Triggered by the backlash over unethical practices of tech organizations, whether inadvertent or intentional, there has been a groundswell of interest in developing code of ethics and national policies. Numerous organizations have

✉ Junaid Qadir, jqadir@qu.edu.qa | [1]Department of Social Sciences and Liberal Arts, Institute of Business Administration (IBA), Karachi, Pakistan. [2]Department of Computer Science, Institute of Business Administration (IBA), Karachi, Pakistan. [3]Department of Computer Science and Engineering, Qatar University, Doha, Qatar.

developing guidelines and principles for developing ethical AI including IEEE's Ethically Aligned Design;[1] Microsoft's AI Principles;[2] DeepMind's Ethics and Society Principles;[3] and Google's AI Principles [36]. There are in fact now a plethora of options for one looking for an AI code of ethics—Mittelstadt noted in 2019 at least 84 public–private initiatives articulating ethical principles that can guide AI [31]. Some ethics certification programs are cropping up for autonomous and intelligent systems such as IEEE's ECPAIS as well as ethical inspection of AI [55].[4]

A recent global survey by Jobin et al. [20] has shown that even though there are certain ethical principles—such as transparency, justice and fairness, non-maleficence, responsibility, and privacy—that are globally considered desirable, there is considerable divergence on how these principles are interpreted and how much each is emphasized depending on the stakeholders' values. Furthermore, most of these ethics codes are non-legislative policy instruments and serve as guidelines or "soft law", which are not legally binding. While all these apparently well-intentioned initiatives aim to ensure that AI technology is used to benefit humanity and lead to their well-being, it is necessary that a critical discourse is initiated to consider the basis upon which a technology may be judged as beneficial and how can it be ensured that the benefits of AI accrue broadly and universally without any undue disadvantage to segments of humanity.

Various prominent modern critics—including Neil Postman, Evgeny Morozov, and even Elon Musk, Bill Gates—have expressed concern over the current trajectory of AI technology and the danger it poses to human civilization and values. This is not unique to AI technology and various philosophers and critics—Martin Heidegger, Jacques Ellul, Herbert Marcuse, and Lewis Mumford—have previously made trenchant critiques of the dangers of technology [51]. From *technological determinism* to *technological lifeworld*, a range of understanding has been developed to calibrate the extent to which technology can dominate its creator [33].

A good degree of optimism, however, still pervades those developing, and funding AI projects. AI developers cite that, "Artificial intelligence constitutes a major form of scientific and technological progress, which can generate considerable social benefits by improving living conditions and health, facilitating justice, creating wealth…" [29]. These reasons are often used to suggest why an optimistic and assuring mood must be exhibited by both states and markets to ensure an unrestricted development of AI. When it is said that, "Intelligent machines are not limited to performing better calculations than human beings; they can also interact with sentient beings, keep them company and take care of them…" [29], it is presumed that all such actions are conducive to increasing social welfare.

But it is here that one must interject and remind oneself that these "benefits" are largely loose presuppositions. The unprecedented changes then being brought upon in the social-moral world by means of AI, which already remains ridden with complexities and issues of inequalities and injustices, ask us to develop a thorough framework of understanding and inspecting such changes before they are taken at face value and termed as benefits. One must also ask who defines what is beneficial and what is not? For instance, can machines taking care of sentient beings be socially beneficial? What kind of a prior socioeconomic-moral landscape would lead to the need for this kind of automated care and in what ways this in turn would affect the development of virtues on an individual level and human well-being on a collective scale? How can we determine whether human or mechanical care is superior since this would depend on the virtues and values we prioritize in our understanding of human well-being?

With growing ethical dilemmas and unforeseen consequences of technology, it is important that we focus on studying the effects of AI technology more holistically in the current global context, which is dominantly neo-liberal democratic and late-capitalist. In this paper, after evaluating the inconsistencies of the current attitudes and models for the social-ethical orienting of the AI, we will discuss and refer to the Islamic guidelines for evaluating AI, highlighting how that has definite and agreed upon sources of guidance, clearly defined set of ethical principles and a strong consensus on the definition of well-being and the conception of a good life.

[1] https://ethicsinaction.ieee.org/.

[2] https://www.microsoft.com/en-us/ai/responsible-ai.

[3] https://deepmind.com/about/ethics-and-society.

[4] IEEE's Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) https://standards.ieee.org/industry-connections/ecpais/.

2 A critical look at the AI revolution: benefits and costs

2.1 Advancements in AI and progress

Far from being confined to a few fringe domains, AI is now well-entrenched in the gamut of civic life affecting the fates of human beings across the world in various high-stake domains such as *education* (deciding who gets or does not get admitted); *banking* (deciding who gets or does not get a bank loan); *policing* (who is released or not released on parole); and *politics* (who gets elected and who does not) [1]. This motivates the need for AI technology should be foolproof and should have gone through a rigorous process of checking for its safety, performance, and robustness. But on the contrary, we now have overwhelming empirical evidence that AI applications are not objective or value neutral [35] or even robust or safe [39]. We now know that notwithstanding the conveniences it creates [5], AI applications disadvantage minorities, engenders inequality, and poses a big challenge to the stability of democratic processes [8, 18, 35].

Can these AI advancements, which enable machines to decide human fate, be labelled as progressive? Rivers claims that progress is progressive, where it is tied to some form of improvement, but not at the expense of some loss [43]. In simpler words, progress is defined as something that provides a gain, but if that gain causes more loss than the gain, we cannot classify the gain as progressive. Even though the technologists are prone to viewing AI's future optimistically, with a new tool or feature being introduced in the name of human 'betterment', a critical evaluation of the numerous ethical and moral questions raised by AI regarding injustice in the world and deterioration in the character and values of human individuals and societies is warranted.

Ironically, one of the main reasons motivating the use of AI applications in preference to human actors, in addition to increased efficiency, was its purported "neutrality" and "objectivity", which was supposed to eliminate such prejudices. The outcome of AI algorithms, however, is dependent on the input, which means if the historical data is biased, the model is likely to be biased as well. Various reported incidents of bias in AI based systems, and the unfair outcomes it can lead to, have shown us that AI models are not neutral or objective, and cannot be assumed to be intrinsically fair, as they only reflect the views of the one who builds them and the data that they are fed [44]. For instance, COMPAS, an AI risk assessment model for predicting defendant reoffending probability used in various US states, was analyzed by researchers, and found to be biased against the black population as many black people were labelled high-risk offenders compared to people of fair complexion who had similar or even worse profiles [2, 24]. Even AI algorithms used for object detection in self-driving cars have been shown to be vulnerable to such a color bias, with higher accuracies reported for detecting individuals with lighter skin tones, implying that black people are more likely to be hit by self-driving cars [11].

Other types of bias, e.g., gender bias, are also rampant. Buolamwini [8] has demonstrated that commercial state-of-the-art AI services are biased against minorities with female minorities suffering with the worst error rates among the various demographic groups. When Amazon attempted to shortlist candidates for recruitment using AI in 2014, it was discovered that their model was likely to prefer male applicants to female applicants with similar credentials [22].

In addition, the inscrutable nature of modern AI tools, which in a black-box fashion provide predictions but without explanations, and the vulnerability of modern AI techniques to adversarial attacks, put into question the trustworthiness of such AI methods and if these methods really deserve the mass adoption in critical systems that we are observing [40].

2.2 Questioning the 'Why' of AI

Another issue we see in this technological era is that we automatically accept the technology for what it is rather than questioning it. Postman describes this as giving technology a 'mythic' status [37]. When technology is granted this status, it is assumed to be accepted as it is, without any attempt to control, question or modify it. The problem isn't the use of AI, but rather the indispensable dependence on technology where it becomes natural to adapt to technological values rather than questioning them. Simply put, when evaluating AI or formulating AI ethics, the emphasis is on 'how', rather on 'why'. We can refer to Heidegger here. For instance, technology promotes ease and convenience, but one must ask why ease and convenience? Why a prolonged life? Why have desires fulfilled? Technology only answers the hows but not the whys. We argue that instead of focusing on just the how, it is more important to focus on the why.

AI comes with the promise of increasing efficiency and unprecedentedly increasing our human control. From smart screens to smart appliances, everything is controlled just with a touch of a button. But this has also resulted in AI applications detaching the humane component from human societies. Why does one need to have control, and to what extent

are human beings supposed to wield such control over their lives? Experiencing losses, bearing pain and suffering, having time alone for introspection, are all part of our human condition, that aid us in moral development, growth, and character-building. Deputing valuable, virtue-instilling tasks, to intelligent machines marks the dichotomy between ease via efficiency and character making via difficulties and challenges.

When there is a debate on the ethics of AI, a common approach is to view certain artifacts in isolation and providing a set of guidelines on how to use them. This approach assumes that technology is value-neutral, and its outcome is dependent on how it is used. It entails that technology in itself does not have a 'will' or an intention, and thereby, it is to be only classified as a tool that is used as the subjects (the humans) desire it to be used [33]. When this premise positing the value neutrality of technology is accepted, the true purpose of why an AI application is being designed escapes scrutiny and is never inquired. This promotes the development of "technology for the sake of technology."

2.3 Ethical dilemmas of AI

It is often thought that AI empowers human beings by providing more autonomy and human control. But at the same time, by posing an ever-increasing number of ethical and moral questions that the current civilization is struggling to answer satisfactorily, technology has also led us to trivialize moral values. Hofmann [19] talks about how technology complicates our existence by introducing questions that challenge our moral and ethical views. His argument is applicable to AI technology as well, where the technology has given birth to various ethical questions and dilemmas that did not exist previously. This has happened in the case of other technologies too, but the nature of these dilemmas has been intensified with AI. For instance, AI models are used to detect fetal heart abnormalities, which can allow for early treatment and might be able to prevent neonatal deaths [42]. However, parents who do not consult the AI model, and have their newborn die raises a moral question, i.e., whether the parents have killed their child by not opting for the technology? Adding on to the diagnostic aids, increases the pressure of decision making on multiple levels, in this case there is an added burden of deciding whether to use the aid and then whether to go for the early treatment. There is a reformulation and reinterpretation of life and death experiences stirred by the availability of such and similar technologies. Moreover, in case of autonomous cars, Valey and Beker talk about testing the AI model used in the cars on real-world roads and raise an ethical question if it is morally acceptable to allow people on public roads to be test subjects for an autonomous car's test run [27]. A similar dilemma can be experienced in the field of medicine and surgery, where the question arises if a future medic robot (well-trained on simulations and in controlled environments) should be tested on real patients [27]. Although experimentation on human subjects has always been subject to ethical debate, there is a new dimension here since other than the status of patient as an experimental subject, the status of the robot as a doctor is also debatable.

2.4 The cost of AI—algorithmic oppression and exploitation

We have until now evaluated the 'potential' benefits of AI technology and 'possible' risks associated with those, we will now shed light on the costs of AI technology that we have paid and will likely keep paying. When talking about the consequences of AI, the conversations mostly revolve around the impact on individuals and industries, and rarely do we talk about the environment. However, research has shown that training AI models, with its massive carbon footprint, is damaging for the environment. A study from researchers at the University of Massachusetts, Amherst, revealed that training of large Natural Language Processing (NLP) human language models can emit more than 600,000 pounds of carbon dioxide, which is *"five times the lifetime emissions of an average American car"* [15, 16]. In terms of energy consumption, the data centers consume approximately 200 terawatt hours (TWh) of energy each year, which is more than the national energy usage of some counties [21].

The costs of AI are not just confined to the environmental costs, there are other costs as well that too often go unnoticed. For instance, consider the cost of injustice that marginalized communities pay. We have highlighted previously the biases against black people and women that AI algorithms can demonstrate. Mohamed et al. identify other instances of biases and algorithmic oppression including facial recognition systems failing to recognize black faces; speech detection models classifying black vernacular as toxic" [30], p. 667). The book "Algorithms of Oppression" highlights similar cases of algorithmic failures specific to women and blacks [34]. One of the examples from the book highlight the sexist nature of Google Search's autosuggestions feature, where when typed "Women should", the search engine suggests: "stay at home", "be controlled", "be in the kitchen" etc. ([34], p. 15). Similarly, in another instance, when searched for "black women", the search results showed the links to pornographic sites ([34], p. 4).

In addition to this, there are also cases where AI technology has been a cause of exploitation of marginalized audiences. For beta-testing[2] of applications, organizations consider countries outside of their own as testing grounds, usually these are low- or middle-income countries or vulnerable populations, since they lack regulations around the data and its usage [30], p. 669). Moreover, a company in China hires low-wage workers to label and tag images that are used to train machine learning algorithms [54]. The same task in Finland is being performed by prisoners [10]. There is little consideration for safety and working conditions for these "ghost workers", who are paid at a very low rate [30], p. 668).

It is important to note that as a victim of exploitation and oppression, a certain segment of the population suffers the loss, while the other has their hopes pinned on gaining what is not even guaranteed.

2.5 AI, and 'good life': a normative approach

Much like the aftermath of the industrial revolution, where the obsession with efficiency that predicated many unethical practices being put into place on factory floors, the evaluation of AI technology by engineers, academics, and policymakers, seems to have run a similar course with an exclusive focus on accuracy, efficiency, and convenience, without much thought devoted to ethical considerations such as fairness, human welfare, and social values. This was the case up until now when more nuanced evaluations are now emerging.

One avenue leading to ethical AI practice is to look for technological solutions to these ethical problems and to strive for better algorithms, and richer and more diverse data sources. Such an approach however is unlikely to work as technological progress has a history of creating newer (often unintended and unanticipated) problems. Furthermore, complex socio-economic problems rarely are solved by purely technological solutions unless they are complemented with appropriate human complements [49, 53]. This realization, along with the manifestation of the various harms and moral dilemmas associated with AI technology, has spurred an entire movement of philosophical study of AI ethics [20].

The meta-ethical question of defining good or good life or virtuous individuals and communities does not engage us here since this paper relies on the revealed sources and the theological cum ethical rules derived from these sources, both deductively and inductively and through other interpretive tools. We undertake the normative assessment to establish a criterion to evaluate the moral status of technology, especially AI. Only by having a normative ethical framework to define what exactly is the value we need AI to seek and what exactly do we require from designers and users to do to realize that value, we can, at a later stage, assess certain practical applications of AI. One approach, the liberal conception, is to assume that there is no universal benchmark for this classification, and the conception of good differs for each individual depending on their beliefs and value system. Without a set of moral principles, the human race risks falling into a nihilistic paradigm, not having firm values to hold onto; values that transcend individual and cultural preferences. Even those who want to question the rights and wrongs of AI, on the basis of higher values and morality, would find serious difficulties doing so from within this framework. The nature of ethical problems ensuing out of AI are global and hence cannot be grounded in the desires and preferences of individuals or specific cultural outlooks. We need human-scale evaluation frameworks for AI technology that has immense potential not only for harm but also for redefining and reframing ontological presuppositions about the nature of human beings and associated concepts.

When there is an absence of a standard to evaluate technology, the technology itself becomes a benchmark of evaluation with no consideration afforded to ethical implications. For instance, the Deep Video Portrait project, showcased during the 2018 SIGGRAPH, is one of several AI projects being developed that aim "*…to demonstrate the capabilities of modern computer vision and computer graphics technology…convey them in an approachable and fun way*" [23]. A project with serious ethical repercussions is being developed with an almost childlike curiosity with the stated goal of enhancing the AI community's understanding of image synthesis, rendering and reconstruction without any consideration for the harms it can engender. This spirit is best captured by the mantra "*Move Fast and Break Things*" [48], once articulated as the motto of Facebook, but which now more or less describes the philosophy of all of Silicon Valley.

2.6 Limitations of consequentialist AI ethics

A common theme that we find underlying most of the problematic AI applications is the implicit assumption that they are beneficial and morally acceptable until any adverse consequences are unveiled. Furthermore, these artifacts

are evaluated in isolation despite the fact that technological artifacts cause widespread changes in terms of altering human thoughts, ideas, processes, values and worldviews, changes that are too abstract yet significant and noticeable enough to be observed and measured. Simply put, the technology creates a huge impact on our *being* and when evaluating a single instance of technology in isolation, these paradigmatic shifts often go neglected.

The *consequentialist* approach to ethics calls for an evaluation based on a utility calculus to capture in advance whether the artifact is beneficial or detrimental. Any such approach not only demands a clear definition and consensus over utility, a project that still remains elusive even for modern advocates of utilitarianism but would also pose problems given that the calculus would always have the imprint of the calculator's bias [51]. A consequentialist approach may then justify widespread changes in terms of altering human thoughts, ideas, processes, values, and worldviews, if the calculator and their calculus give this a high score. Measuring all the consequences together for all future possibilities is not pragmatically viable. Value-alignment theorists aim to derive the desired "values" to which machines should align their behavior. These values might reveal an individual's or a community's preferences, but not necessarily indicate what is right, good, just, or appropriate [26].

Further still, the changes that technology, especially AI technology, brings with itself are often opaque and irreversible. Vallor [51] describes how the *acute technosocial opacity* of modern technology makes it "difficult to identify, seek, and secure the ultimate goal of ethics—a life worth choosing". In particular, it is hard to predict how an extremely powerful technology will impact the human socio-economic ecosystem as it co-evolves with human beings and society over time. There are several AI technologies whose harms have now fully manifested, and it is now too late to undo them. The use of AI to develop ultra-personalized news feed algorithms by social networking sites such as Facebook poses serious threats to the integrity of democracy processes such as elections [50]. The use of deep learning based AI models has resulted in the creation of fake images and videos resulting in cases of defaming, manipulation, and invasion of human privacy. To summarize, the outcomes of AI technology are often opaque but also drastic, irrevocable, and widespread. The design of AI artefacts demands careful proactive thought as we cannot afford to wait for their consequence to reveal before we can evaluate the moral value of the artifact.

2.7 Virtue based AI ethics

Researchers are now turning to virtue-based ethics as a viable framework for responding to the various dilemmas that the AI era proposes [51]. If we look at the virtue ethics' conception of human character development, it is human beings who need to strive to rise above their biases in the quest for living by truth and justice. Virtuous human societies do not just need efficient, unbiased court trials. Virtuous human societies more than anything else need better human beings in pursuit of and exhibiting higher moral ideals.

We then understand that our position with AI should be uncertain at best. A careful analysis of AI demands that a standard framework should be set for its evaluation. In continental philosophy, Kant's *deontology* and Bentham and Mill's *utilitarianism* remain the two most widely prescribed normative approaches to resolve ethical dilemmas. One can argue that their extension can also be used to analyze larger societal repercussions of AI. However, both the approaches remain troubled by their own unique issues.

Consequently, a renewed fervor has lately been exhibited in seeing virtue ethics as an answer to develop a more nuanced understanding of the ethical and social dilemmas of the 21st century. A renewed interest in Aristotelian, Confucist, and Buddhist versions of virtue ethics has commenced [51]. Nonetheless, with the renewed interest in virtue ethics, efforts have been made to come up with a comprehensive virtue-based ethics system for the ethical development of AI.

Of these efforts, Shannon Vallor's book "*Technology and the Virtues*" [51] is the most prominent and aligned thematically with the questions that we have highlighted. She states, "*Yet it remains the case that very often, the answers for which questions about emerging technology beg are simply not of the 'yes/no' or 'right/wrong' sort.*

Instead, they are questions of this sort: 'How might interacting with social robots help, hurt, or change us?'…" Her solution is the development of "technomoral virtues" that allow us to have very certain and carefully defined universal principles that help guide our attitude with different changes being brought upon by the AI. Vallor is quick in understanding the potential of conflict that can arise if a relativist approach is taken with ethics. Yet, with forces of globalization and late capitalism, interacting with each other in a dynamic political context, it is extremely hard for States, and even more, trans-national corporations, to agree and adhere to a single set of rules. Even when a single set of rules is defined, one can understand how conflicts would quickly emerge as different cultures and people from around the globe begin interacting with them. Perhaps some would have prioritized some of their goals over others that is the economy over the environment, among other trade-offs.

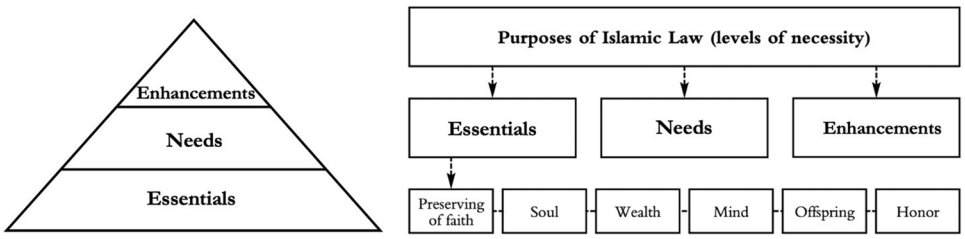

Fig. 1 Specifying the Levels of Necessity in *Maqāṣid* discourse () adapted from [6]

However, it is precisely this predicament that Vallor tries to overcome by suggesting different technomoral virtues that allow us to cultivate a "*technomoral wisdom*". Realizing the potential differences between the different classical virtue ethics systems, Vallor's core argument focuses on the central virtues and qualities preached by each system that are shared by all systems. In this way, as our interdependence in the techno-social world increases through increased globalization, and our challenges and concerns over AI become universal, Vallor's technomoral virtues become an aid in action-guidance that focus on the roots of virtues rather than focusing on their ultimate deliverance. She elaborates that in these words, "*Our present moral context, the one to which our virtues must be more effectively adapted, is one of increasingly rapid, transformative, global, unpredictable, and interdependent technosocial change.*"

In this way, by focusing on particular virtues, Vallor's thesis emphasizes that "a set of behavioral, cognitive, perceptual, and affective habits" are "needed to cultivate oneself in any moral world." She highlights 12 technomoral virtues we must focus on including honesty, self-control, and humility. To elaborate further, she defines technomoral virtue of honesty as "…an exemplary respect for truth, along with the practical expertise to express that respect appropriately in technosocial contexts." Whilst Vallor's thesis moves in the right direction, it fails to develop a holistic understanding of human nature that should share the center stage with her technomoral virtues. Her thesis, although normative and thus, addressing the question of what ought to be, does not bring into its fold the idea that the post-modern individual still remains objectified and constantly trapped in the chasms of the late-capitalist neo-liberal world and how that interferes with the cultivation of techno-moral virtues. The question of ethics is inherently always connected to a metaphysics through which it drives its system. Vallor, in trying to ensure her virtue ethics does not comment on the essential nature of humans, thereby allowing her system to be flexible and universal, does not say much on this inherent connection [14].

Any virtue ethics system should prescribe certain virtue to cultivate a certain conception of a harmonious society as its goal. We need a normative ethics that has two poles: one is an agreed upon set of values that has the potential to be universally applicable to humanity at large, despite allowing for some pluralistic cultural nuances, the other is a set of virtues that can be cultivated by individuals, which are informed by the normative framework and in turn help establish its ethical objectives values. The relationship between the normative set of values or objectives and the cultivation of virtues would be multidimensional. One goal for cultivating virtues and excellence in character would be striving for moral and spiritual excellence. These virtues would help educate the minds and hearts in ways that would allow them to transcend the instrumental and calculative aspects of technology and think holistically so as to reach collective benefit and well-being, in alignment with the normative objectives. Having a community of virtuous individuals, would make the need for many AI applications, especially those that are aimed to substitute human services, redundant, since in order to improve morally, there will be strong social bond, mutual cooperation and compassion within the communities. The normative framework would allow an evaluation, arbitration, and determination of a set of collective values that are at the same time objective, to let the individual virtues grow in directions that foster those values. The virtues, once firmly instilled, should allow technomoral choices to be made as individuals users and collective units of designers, developers, academics, and policymakers. Without the normative framework of values, that acts as a regulatory framework, the virtues remain too individual to be able to consciously contribute toward the creation of right AI policies for creating harmonious societies. We are in dire need of intra and inter harmonious communities, globally, working toward common good and well-being, while benefiting from their cultural particularities. Without the individual development of virtues, the normative policies could not be properly understood, in their multimodality and hence not actionable in collective decision making.

2.8 Objectives and purpose (*Maqāṣid*) of the Islamic law (*Sharʿīah*)

The Islamic tradition, followed globally by more than 2 billion people, has a rich ethical tradition spanning more than 1400 years, that people in the Muslim world are closely tied to. The Islamic tradition is comprehensive and encompasses a

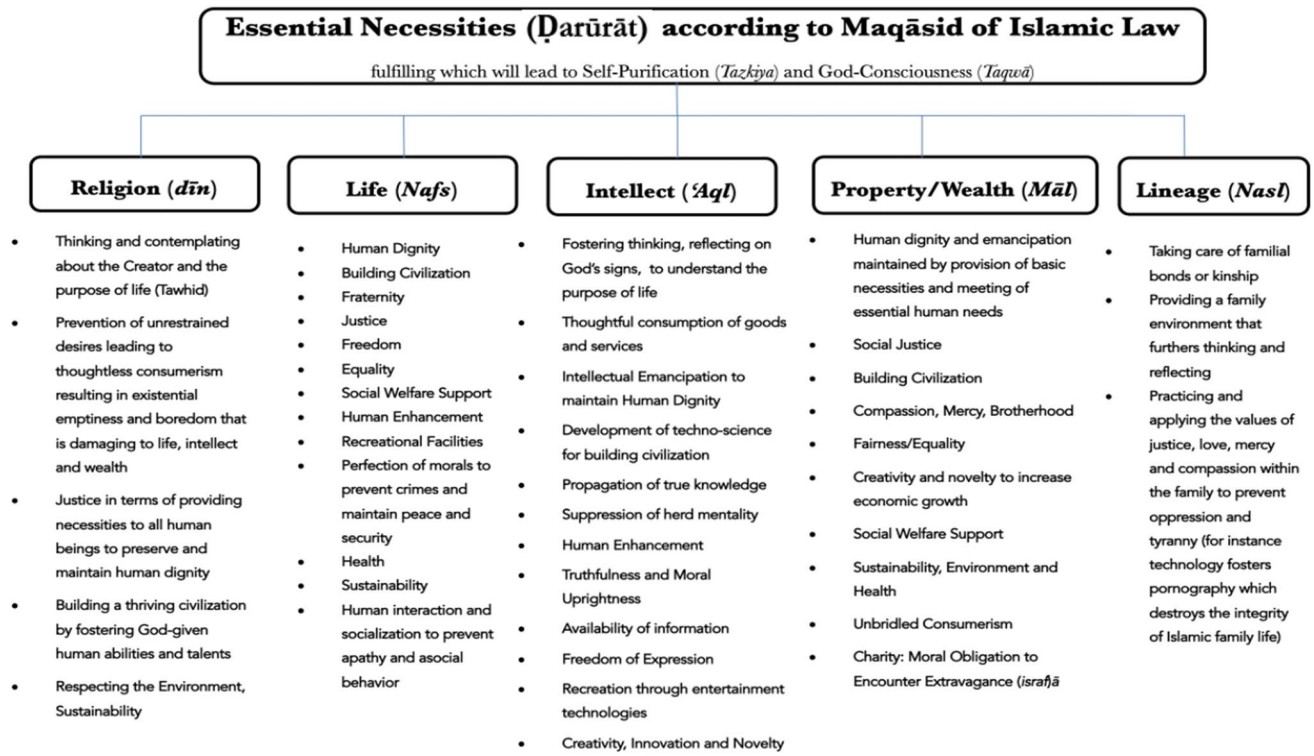

Fig. 2 Essential Necessities according to *Maqāṣid* of Islamic Law

spectrum of solutions ranging from concrete legal instruments and incentives for the community as well as for individuals. Islamic scholars have extrapolated in the light of Qurʾān and *Sunnah* that the *Sharʿīah* has certain higher objectives and purposes (*Maqāṣid*). The *Maqāṣid* theory as developed historically acts as an ethical compass that allows Muslim communities to live out all dimensions of their life in accordance with the *Sharʿīah*. By applying the *Maqāṣid* approach, we can discover ethical principles for all situations faced by human beings.

The *Maqāṣid* follow a hierarchy in terms of prioritization of the necessities, needs and enhancements as can be seen in Fig. 1. *Essentials* (*ḍarūrāt*) refer to absolute necessities; *Needs* (*ḥājiyyāt*) are necessities to a lesser extent; while *Enhancements* (*taḥsīniyyāt*) are dispensable needs that are desirable nonetheless for beautifying/facilitating purposes. At the level of *Essentials*, *Sharʿīah* necessitates that the five essential objectives—including religion (*dīn*), life (*nafs*), progeny (*nasl*), property or wealth (*māl*) and intellect (*ʿaql*)—should be preserved. This is the traditional *Maqāṣid* classification (Fig. 2) given among others by al-Ghazzāli. Whatever has the potential to cause harm to any of the five objectives is strongly prohibited.

The *Essentials* are critical for the preservation and sustenance of the *Maqāṣid*; while *Needs* and *Necessities* support and complement the *Essentials*; and finally, *Enhancements* complement the *Essentials*, while making improvements to the five objectives. The *Essential* elements have priority over *Need* elements, which are to be prioritized over *Enhancements*. It is also important to note that the three categories are not absolute and vary depending on the circumstances of individuals and societies; however, the good of the community always has priority over the good of an individual.

In addition, *Maqāṣid* can also be classified according to the scope of the people included in purposes and the level of universality of the purposes [6]. For example, a famous Islamic scholar Ibn Ashur gave the Maqāṣid that are concerned with the community (*ummah*) priority over those that are concerned only with individuals. Some scholars such as Muhammad al-Ghazali have included *justice* and *freedom* in *Maqāṣid* at the level of necessities [6]. Some scholars have added the preservation of honor to the five popularly known necessities [6].

2.9 *Maqāṣid* based ethics for technology

The objectives at the level of essentials (necessity) [*ḍarūrāt*] are most important. Raquib [41] has proposed an Islamic *Maqāṣid* based ethical framework for technology that suggests a holistic analysis, keeping in mind that contemporary, late-modern technology, when seen from a broader lens, reshapes cultures, worldviews, ideas and even, redefines harms and benefits. She has derived various essentials/necessities (*ḍarūrāt*) of the human society from the Islamic tradition and classified it under the five central objectives developed by the classical Muslim scholars illustrated in Fig. 2. The content under each objective in Fig. 2 above, has been proposed by the author and is not an agreed upon list.

2.9.1 Harms and benefits

The *Maqāṣid* approach is closely linked with the concept of *Maṣlaḥah* (equivalent to 'Benefit' or 'Human Good') since one needs to have a thorough understanding of Human Good to evaluate an act using the Objectives of *Sharʿīah*. In cases where no direct textual reference from Qurʾān or Sunnah is available to sanction the desirability or undesirability of the situation, it should be judged by evaluating the Human Good (*Maṣlaḥah*) and Harm (*Mafsadah*). It is important to note that the Islamic idea of Benefit covers both individual and societal welfare in all their dimensions, therefore, when evaluating the human good and harm, both individual and social contexts are considered in the light of customs and known practices in any given situation. Moreover, the priority is given to collective good over individual good, and prevention of collective harm has priority over prevention of individual harm and acquisition of collective good. To re-emphasize, if attaining good for an individual causes unintentional harm or even has a probability to cause harm to others, it would then be considered as prohibited. At this point, if we take a step back and review harms and benefits ensuing from the AI technologies, we observe that AI is assumed to result in many benefits and a few harms. We still do not get a clear answer on benefits and harms for whom? Is it sufficient if AI is only beneficial to the designers or developers or a certain sector of the society? For instance, if we consider the example of industrial robots, those may have benefitted industries by increasing efficiency that is also cheap, but have they also benefited the society as a whole, or have they resulted in mass unemployment?

2.9.2 Categorizing the 'good'

Modern technology has blurred the lines between needs, wants and luxuries. The *Maqāṣid* theory makes this distinction by classifying 'benefits' or 'good', obtained from preserving the five objectives, in three categories: (1) essentials; (2) needs and necessities; and (3) enhancements.

For instance, in the case of AI, the use of AI for diagnosing diseases has been quite prevalent [28]. Automated diagnosis makes the process more efficient, that is, diseases are diagnosed more quickly. The efficiency element is a benefit that 'improves' the process of saving lives, which is why we can categorize it as an *Enhancement*. However, in a scenario where a certain diagnosis goes wrong and an individual's life is put at risk, then this risk causes harm to the *Essential* of the objective of Life. In simpler words, trying to save someone's life is critical to preserving the objectives of life, whereas being efficient at saving someone's life is an improvement. Therefore, applying the Islamic objectives ethics, the use of AI for disease diagnosis simply for the sake of scaling up efficiency would be avoided if it risks human lives more significantly than what would have been the case with a human doctor.

2.9.3 Blocking the means (*Ṣadd al-Ḍharai'*)

Another important concept in the Objectives approach is of blocking the means to unlawful or unethical ends. The concept allows for the process of identifying and eliminating any means that can possibly lead to corruption. However, unlike the traditional consequentialist approach, the end results are to be 'judged in relation to the religio-ethical standard of the preservation of the five objectives', which means the consequences are not evaluated merely on the basis of an individual's idea of harm and benefit, but rather on a well-defined set of ethics that protect the objectives. For instance, an article published in Harvard Journal of Law and Technology discusses an application of machine learning in criminal justice, where recidivism-risk score is used as an input to make sentencing decisions [12]. In this case, there is a risk of greater harm if an individual is wrongly sentenced. The principle of *Ṣadd al-Ḍharaiʿ* (blocking of means), will prohibit

such a system since the end could result in compromising an individual's life, which is one of the five objectives that needs to be preserved.

2.10 Islamic Maqasid based virtue ethics for AI approach

What remains absent from the recent efforts on virtue ethics is the study of the rich body of literature on virtue ethics produced in the medieval and premodern Islamic world [9]. Scholars from the Islamic tradition including intellectual giants such as al-Ghazzāli [3], al-Shāṭibī [4], and Shāh Walīullah [46], had comprehensive and nuanced philosophical thinking, particularly in the areas of human psychology, divine providence, morality, spirituality, and virtue. The Islamic idea of virtue is infused throughout the ethical-spiritual system within the broader framework of Islamic Sharia which is the embodiment of the Islamic worldview. The objective or *Maqāṣid* based ethics rooted in the Islamic metaphysics articulates that the human beings, the universe, and the creation of this world, all have their purposes and objectives (i.e., they have *maqāsid*). This distinguishes *Maqāṣid* based ethics and enables it to put standards of virtue based on Islamic guidance into practice. The Islamic virtue ethics, as part of *Maqāṣid* discourse, needs to be formally developed. The methodology proposed below aims to provide a virtue theory grounded in a foundational yet broad normative framework, rationally derived from an inductive reading of the Islamic scriptural sources.

When we conceptualize virtue-based ethics within the Islamic tradition, we mean an ethical framework to settle the hierarchy of values within the social realm that is then concretized by a societal mechanism of inculcating virtues within individuals resulting in virtuous collectivities, who can uphold the Islamic ethical objectives in their individual and collective lives. We, in this paper, frame the *Maqāṣid al-Sharʿīah* (objectives of the Islamic *Sharʿīah*) as the normative ethical framework to define what exactly are the values that AI technology needs to seek, in the social realm, that the designers, developers, engineers, policy makers and users can then translate into actions through their virtuous dispositions. The cultivation of virtues, although reliant on an experiential learning through life challenges and opportunities, would in this instance be informed by the *Maqāṣid* ethics framework and the general virtues, agreed upon by the major religious and secular ethical traditions, could be integrated within the broad umbrella of the five or six *maqāṣid* (objectives) (shown in Fig. 2), specifically under the religious objective of *tazkiyah* (self-purification).

The technomoral virtues required for technomoral wisdom, referred to in the previous section and having a global appeal, can be appropriated from within the *Maqāṣid* framework, so that the ethical values in the form of *maqāṣid* (objectives) help in the virtuous character making of the societal members. These virtues would then enable them to build a technomoral society as they live by those values and integrate those in their individual and collective choices. The objectives and the general virtues overlap in this proposed scheme with the technomoral virtues becoming a subset of both. Abdur Rahmane Taha emphasizes the practical dimension that is the praxis informed by deep thinking or the theory and in turn informing the theory. The expression of self-worship and narcissism-of both the designers and users of technology—most evident in social media technologies, will find further reification in AI applications meant for human mastery and control. Paradoxically, modern human subjects, at the pinnacle of self-love, surrender themselves to the techno-market forces, under the illusion of sovereignty, mastery, and control. He elaborates upon the kind of subjectivity to be formed that would allow the subjects to exercise enhanced moral-practical reasoning, so the virtues inform their moral decisions in the capacity of designers, engineers, and users of technology [17]: pp. 163–64, [47]: p. 62).

When it comes to evaluating the AI technology, a common approach is to view certain technologies in isolation and analyze their harms and benefits. The Islamic ethics outlines clear guidelines for all walks of life and is based on the original sources of Islamic legislation, i.e., (1) the Holy Qurʾān; and (2) the *Sunnah* or *Hadīth*, which refers to the traditions of sayings, approvals, and actions reported from the Prophet Muhammad (peace be upon him). The Islamic ethics endorses all dimensions of individual and social life, everything that is experienced either inwardly or outwardly falls within the jurisdiction of the Islamic Law (*Sharʿīah*). The scriptural sources are not univocal and there are areas of non-consensus, still since the source of Islamic ethics and law is divine, it gives way to a divine unity among Muslims where they can always refer to the Scripture to build consensus, specifically in unprecedented situations.

3 Discussion

3.1 Challenges in operationalizing AI ethics

Most contemporary AI development, globally, takes place mostly in pursuit of economic incentives in partial, if not complete, isolation from fundamental ethical discourses. As a result, ethical questions are ignored and sidelined until real economic (or material) costs are attached to the equation. Beyond initiating AI ethics conversations by questioning the larger schema that orders our world, there is a more pragmatic hurdle that AI developers face. The unavailability of any complete pre-existing tool or methodology that can be readily used by AI developers to test the ethical nature of their project is a bitter ground reality. Although several documents addressing ethics in AI have been produced—including Montreal Declaration,[5] OECD Council Recommendation on Artificial Intelligence[6]—most documents contain only abstract normative principles [20]. These documents have still not been translated into a tool or methodology that can be used by AI developers, universally, to check the ethical nature of their projects [31].

Nonetheless, Jobin et al. [20] have noted, after review of 84 such documents, that more than half of them feature the shared themes of transparency, justice and fairness, non-maleficence, responsibility, and privacy. In addition, Floridi et al. [13] have reported that most of these ethical principles converge around five principal virtues: (1) transparency, (2) justice and fairness, (3) non-maleficence, (4) responsibility, and (5) privacy. From this fragile consensus we understand that a mutual foundation on AI ethics can be built to evaluate deliverables and communicate expectations [32]. The consensus on these ethical principles, although universal only to a limited degree, is something that is reiterated across governmental bodies, academia, and tech-industry [32]. However, the question of how to translate the principles into practice still remains unanswered.

The said documents, even if they converge upon certain ethical principles, exclusively create normative ethical discourses around the development of AI. These discourses are essentially treated as mere guidelines as the independent development of these guidelines into tools or methods, by any enterprise, is very challenging and are deemed by some as overheads. These overheads are further seen undesirable as they offer no short-term commercial incentives whereby the advantage of developing an ethically aligned AI remains unclear [32]. To respond to this dilemma, Morley et al. use the idea of fragile consensus, as stated above, to develop a universally applicable Applied AI Ethics typology around the five ethical principles. However, their efforts are met with complications like uneven distribution of tools across different stages of project development to ethically assess the AI project whereby evaluation at certain stages for certain ethical principles is harder. Moreover, their typology suggests a neat distinction between the different stages of technological innovation, which may not be the case.

3.2 Need for holistic value and purpose based analysis of AI

Since the early 20th century, many paradigms, including positivism, have seen their entry and exit in the larger academia. The scientistic belief that the ultimate reality of the universe and its meaning can only ever be seen through the scientific lens reduces all human experience to just an array of scientific phenomena. But the purported "value neutrality" of science and technology, its promise of "absolute objectivity", and assurance of a suffering-free world as a reward to believe in its forces, has never seen itself really be questioned or revisited axiologically. Under this broader philosophical discourse then, we can observe why any attempt to evaluate AI would only evaluate its external, objective embodiments (i.e., performance, speed, and efficiency).

Yet any such template would fail to truly evaluate AI's broader, more intrinsic, implications on human societies for these call upon value judgements to be made which, if given, would be antithetical to the very nature of the technological template of evaluation that claims value neutrality. The evaluation of AI technologies thus must take place using a framework that is other than that which is technological/scientific. Instead, a framework that embodies a more comprehensive and holistic understanding of human nature and experience—such as the techno-moral framework of virtues and habits that this paper develops through the Islamic perspective—must be used for holistically analyzing AI.

The charter of human well-being, when erroneously defined in scientific and technological terms, is often seen captured through numbers as attempted in neoliberal economics, among other soft sciences. By trying to conclusively

[5] https://www.montrealdeclaration-responsibleai.com/the-declaration.
[6] https://www.oecd.org/going-digital/ai/principles.

argue the human condition in different countries using various, but almost always reductive, figures such as the gross domestic product (GDP), secular-modernists failed to truly capture the human experience for they were only able to inspect a singular dimension. We must be wary of opting for a similar approach for AI noting that the very same undertaking eventually became part of the 1960's developmental discourse—consequential to the fragmented evaluation of the developing world that led many countries to pursue secularist neo-liberal economic goals that were detrimental to their populaces and only exacerbated scars left by colonization whilst setting the stage for neocolonialism.

This then points us to another problematic aspect of AI which is its potential to quickly spread across the globe becoming a paradigm on its own that affects the human race changing the very scheme of economic and social structures governing the world currently to make them even more inequitable. This can be loosely, if not completely, inspected through different phenomena including the digital divide, which provides an insight into how newly arising techno-structural inequalities may look like. The progress and pervasiveness of AI may aggravate phenomena like the digital divide and other sorts of inequalities resulting in more harm (e.g., job losses) than good (principally, increased efficiency). Thus AI, even though it is viewed often as a value neutral tool, has the potential to act as a catalyst for growing techno-capitalism that reinforces the inequalities of late-capitalism and the emerging surveillance capitalism.

The evaluation of AI and the subsequent techno-moral virtues and habits then must come from beyond the confines of Silicon Valley technologists who celebrate anti and post-humanistic libertarian capitalist values that predominantly rest on a continuance of late-capitalism and consumerism. Well-being and good life, within these trenches is defined as a blind pursuance of efficiency and commodification [14]. Instead, to truly evaluate AI, we must readjust our understanding of well-being and good life as offered by various ethical approaches. Whilst the former would evaluate AI based on techno-social norms that currently govern Silicon Valley—gross consumerism and surveillance capitalism [56], the latter would examine AI more broadly while considering the acute techno-social opacity we face [51]. Knowing the problems that feature themselves when well-being and ethics are understood and practiced through a utilitarian or deontological approach, two of the most popularly sought and studied ethical paradigms, a virtue ethics approach, as one presented in Islam, would seem a better alternative.

Auda [7] describes why it is important for Maqasid scholarship to avoid the traps of partialism and apologism (among other things). It is important to be purpose and objectives driven and this allows us to steer the present appropriately to the future. Thus, Maqasid based Islamic AI ethics should be holistic, comprehensive, dynamic, and future-oriented. Revelation anchors the Maqasid scholarship with a stable divinely-defined worldview, to which it can consistently return [7]. It is important to note that the purposes and objectives are multi-faceted and interweaved and active at the same time. The objectives cannot be considered in isolation. Thus, the objectives of preserving faith, life, mind, progeny, wealth, and dignity in the classic Maqasid framework should not be considered in isolation as if there is no interaction between them since such partialistic analysis can result in judgements that defy the established Islamic universal principles. Auda [7] highlights how a systems-focused holistic Maqasid methodology is needed to offset such partiality thinking that can emerge in human thought in general with Islamic Scholarship being no exception.

3.3 Salient points of agreement among Islamic AI ethics community

In December 2021, two of the authors of this paper (the first and last) were part of a team that convened the First International Conference on Islamic Ethics and Artificial Intelligence in Lahore, Pakistan.[7] The conference was a hybrid online and in-person conference with participation from various Islamic scholars, Muslim AI professionals, AI ethicists, and experts in policy and design. In what follows, we report the salient agreements that emerged after two days of discussion between the participants in the First International Conference on Islamic Ethics and AI organized in Lahore, Pakistan in December 2021.

1. AI has the potential for use that greatly benefits humanity and also potential for great harm.

[7] The First International Conference on Islamic Ethics and AI. This conference was organized as part of the project "*Culturally-informed pro-social AI regulation and persuasion framework for Pakistan and the Muslim world*" funded by Facebook Research Ethics in AI Asia–Pacific, whose support as an unrestricted gift is gratefully acknowledged. We note that the sponsors did not set or influence the agenda of the conference in any way; the organizers have tried their best to engage relevant stakeholders and provide them an opportunity to provide independent critical input. The program schedule and material are available at https://www.islamicaiethics.info/.

2. AI technology acts as an amplifying force. AI can be used for doing social good or social evil much more forcefully than what is possible without AI.
3. Problems wrought by technology will not be solved by technology alone. Ultimately people and their habits and mindset have to be transformed according to the higher values sought for the society.
4. Benefits and harms, according to the Islamic perspective, include, but are not limited to, material and worldly concerns.
5. In an Islamic worldview, the human interests of religion, life, intellect, wealth, lineage, and dignity demand systemic preservation and protection through ethics, policies, and law.
6. When venturing into new domains where there is potential for both benefit and harm, warding off harm has priority over gaining potential benefits, and harms and/or benefits known with certainty or high probability are prioritized over posited benefits or harms.
7. AI should not be used to promote or aid injustice (*ẓūlm*) in any of its forms.
8. AI should not be considered parochially with a narrow focus on the vested self-interests of any individual person or corporation or country.
9. The big tech corporations, many of which have global users that outnumber large nations, are managed according to their own commercial interests. There must be national and international mechanisms, institutions, and agreements, that help ensure that technology is used in a way that benefits all of humanity.
10. Ethical use of AI cannot be realized just through principles. An entire system of complementary subsystems (moral, ethical, educational, economic, legal) promoting human-beneficial AI is needed.
11. The effects of AI should be studied holistically with a universal outlook (*ray al-kulli*) focusing on individuals, communities, and the environment. The focus on the universal outlook can provide a common ethical platform that allows alignment with other religious and ethical philosophies that emphasize human good and well-being.
12. There is a need for a system that thinks for all humanity. Islam can provide this inclusive system that caters to and promotes the welfare of all humanity.

3.4 How Maqāṣid based Islamic virtue ethics can contribute?

Though the typology proposed by Morley et al. [32] fails to address certain problems, it provides a usable framework for how Islamic virtue-based ethics may be instantiated. We observe laws and regulations being composed following the typology proposed by Morley et al. can principally, and partially, be juxtaposed to the Islamic virtue-based ethics system. For instance, the principle of non-maleficence in many ways reflects the concept of *Ṣadd al-Ḍharai'*, with both emphasizing the prevention of harm. This similarity then suggests that there exists a great body of documents and literature that embodies the Islamic virtue-based ethics system. The central problem that arises in the schema of Morley et al. is not that of its logical incoherence, but of the structural limitations the tech industry imposes on any discourse on ethics. It is precisely this why they suggest short-termism to be a limitation that affects their typology. In retrospect, if the five ethical principles are enriched and complemented by the three-level *Maqāṣid* hierarchy—i.e., *Essentials*, *Needs* and *Necessities*; *Enhancements*—to evaluate the project at its different stages, a schema could be fashioned that would have the potential to be just robust enough to deal with the nuances their typology faces.

The *Maqāṣid* theory approach would necessitate a greater need to develop usable tools and methods to evaluate AI at different stages. Additionally, where the scope of the said typology was limited to the individual AI projects, the *Maqāṣid* theory approach would apply more broadly and will necessitate fundamental changes in the very lens in which AI is conceived. Questioning the why of the AI project, and how its benefits would become permanent, and would have to be evaluated at each stage of the project. Those who are pro self-driven automobiles, bring the argument that the rate of accidents is quite high with human drivers and this can be reduced by automated cars.[8] In addition to not knowing with certainty if self-driven cars will cause less or more accidents, the ethical goal here does not only pertain to minimizing accidents, but also to inculcate responsibility and empathy in drivers who need to drive which is an equal social good for the healthy functioning of any society. If we do not morally train these individuals or drivers, then the automated cars might prevent road accident deaths but not the deaths these same individuals cause due to their intemperance or rage. The objectives-focused Islamic virtue ethics in practice would enable this foresight. By observing, and utilizing, the typology suggested by Morley et al., we realize that the *Maqāṣid* theory approach has a radically larger scope. Unlike most ethical discourses surrounding AI that remained focused on individual dimensions of AI, the Islamic virtue-based ethics system tries to mitigate the structures of inequality that remain pervasive in the tech industry. Under this system,

ethics is given a greater central stage in the individual's private life whereby an individual's ethical choices shape the society they live in; and in return, the society shapes the individual's ethical choices. This continuous feedback mechanism ensures that the private individual is intimately tied to the social whereby the individual becomes the social. This then ensures that any structures of inequality, that are usually the limitations we encounter with suggesting an ethics system for AI, are quickly identified and rectified.

In addition to this, under this system, the very conception of what an incentive is, changes. We understand that the current schema that orders the world requires any action, be it a social good or bad, to be motivated essentially by economic (or material) incentives. Resultantly many AI applications are developed without properly being evaluated for all their respective repercussions. However, under the *Maqāṣidic* Virtue theory, institutions, and corporations, would not evaluate their laws or projects, respectively, on the basis of the economic incentive they create, but would have to be judged to see if it harms any of the five objectives that *Sharʿīah* has laid down. For instance, an AI that promises to preserve life (*nafs*), such as predictive policing aiming at protecting individuals, would also be evaluated for the potential harm it can cause to life (*nafs*), and check if predictive policing is leading to more people of color being arrested, jailed, or physically harmed [45]. The very outlook then with which AI is approached would be fundamentally different under such a system.

4 Conclusions

In this paper, we have discussed how the ethical bankruptcy of the contemporary AI approaches, which proceed on a vision of innovation for the sake of novelty, profit, and economic growth, grounded in the contemporary neo-liberal late capitalist post-modern socio-economic thinking, has resulted in various ethically dire consequences. We describe how despite the resurgence of interest in virtue-based ethics for AI, previous works have not explored the rich Islamic tradition for insights. This paper presents a holistic, Islamic virtue-based ethics framework for AI that does not suffer from the problems plaguing consequentialist and deontological ethics, which both lack intention of virtue and goodness. This framework instead aims at the cultivation of virtuous individuals and communities that can work together to live a teleological purpose-driven life, where there is a shared conception of good.

Our Islamic virtue-based AI ethics framework builds upon the existing ethical work taking an objectives (*Maqāṣid*)-focused approach. In this framework it is not just a set of desired or undesired consequences that would determine whether an AI application should be commercialized or not. We discuss how various ethical and religious traditions across time and space not only have similar critical stance toward the negative consequences of technologies but have lots of commonalities in their virtue ethics formulations, which can act as a potential template for developing a universally shared ethical standard for the development of AI. We believe an Islamic AI ethical framework is more likely to be adopted by the global Muslim population since a moral and legal ethical code rooted in the local culture, tradition, and values has a greater chance of being accepted rather than a code perceived to be foreign and alien. We also posit that incorporating ideas and engaging with the Islamic virtue-based AI framework will enrich the global AI ethics discourse and provide a basis for dialogue.

Acknowledgements The authors would like to thank the presenters at the First International Conference on Islamic Ethics and AI for their input. In particular, the authors would like to thank Dr. Aasim Padela and Dr. Ezieddin Elmahjub for their help in conceptually improving this paper and Dr. Padela, Dr. Shahzeb Khan, and Dr. Jasser Auda for providing input on the salient points of agreement among speakers noted in this paper.

Author contributions AR and JQ conceptualized the paper together. AR wrote the first draft along with authors BC and TZ. JQ substantially edited the paper, added new sections, added new figures, rewrote the conclusion, and reorganized the paper as needed. All the authors reviewed and contributed to the manuscript. All the authors read and approved the final manuscript.

Funding The authors acknowledge funding from Facebook Research as part of the Ethics in AI Research Initiative for the Asia Pacific for their project "Culturally-informed pro-social AI regulation and persuasion framework for Pakistan and the Muslim world". The research for this work was performed, and the research grant funding this project was submitted and awarded, while the last author was at the Department of Electrical Engineering, Information Technology University (ITU), Lahore, Pakistan.

Declarations

Competing interests The authors declare no competing interests.

References

1. Ahmad K, Maabreh M, Ghaly M, Khan K, Qadir J, Al-Fuqaha A. Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. Comput. Sci. Rev. 2022. https://doi.org/10.1016/j.cosrev.2021.100452.
2. Alex. Racial bias and gender bias examples in AI systems. San Francisco: Medium; 2018.
3. Ghazzālī A. Essential Ihya''Ulum Al-Din—the revival of the religious sciences. (F. Karim, Trans.), vol. 4. Iran: Islamic Book Trust; 2019.
4. Aḥmad A-R. Imam al-Shatibi's theory of the higher objectives and intents of Islamic law. (N. Roberts, Trans.). Herndon: International Institute of Islamic Thought; 2013.
5. Ali A, Qadir J, ur Rasool R, Sathiaseelan A, Zwitter A, Crowcroft J. Big data for development: applications and techniques. Big Data Anal. 2016;1(1):1–24.
6. Auda J. Maqāṣid al-Sharʿīah as philosophy of Islamic law: a systems approach. Herndon: International Institute of Islamic Thought (IIIT); 2008.
7. Auda J. Re-envisioning Islamic scholarship: Maqasid methodology as a new approach. Brynmill: Claritas Books; 2022.
8. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification Conference on fairness, accountability and transparency. New York: PMLR; 2018. p. 77–91.
9. Bucar EM. Islamic virtue ethics The Oxford Handbook of Virtue. Oxford: Oxford University Press; 2018. p. 206–23.
10. Chen A. Inmates in Finland are training AI as part of prison labor. Washington: The Verge; 2019.
11. Cuthbertson A. Self-driving cars more likely to drive into black people, study claims. Kensington: The Independent; 2019.
12. Donohue ME. A replacement for Justitia's Scales: machine learning's role in sentencing. Harv J Law Technol. 2019;32(2):657–78.
13. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Vayena E. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach. 2018;28(4):689–707.
14. Gupta A, et al. The state of AI ethics. Montreal: Montreal AI Ethics Institute (MAIEI); 2021.
15. Hao K. China has started a grand experiment in AI education. It could reshape how the world learns. Cambridge: MIT Technology Review; 2020.
16. Hao K. Training a single AI model can emit as much carbon as five cars in their lifetimes. Cambridge: MIT Technology Review; 2020.
17. Hallaq W. Reforming modernity: ethics and the new human in the philosophy of Abdurrahman Taha. New York: Columbia University Press; 2019.
18. Helbing D, Frey BS, Gigerenzer G, Hafen E, Hagner M, Hofstetter Y, Zwitter A. Will democracy survive big data and artificial intelligence? In: Towards digital enlightenment. Cham: Springer; 2019. p. 73–98.
19. Hofmann B. Ethical challenges with welfare technology: a review of the literature. Sci Eng Ethics. 2013;19(2):389–406.
20. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389–99. https://doi.org/10.1038/s42256-019-0088-2.
21. Jones N. How to stop data centres from gobbling up the world's electricity. Berlin: Nature News; 2018.
22. Kantarci A. Bias in AI: what it is, types & examples, how & tools to fix it. Estonia: AIMultiple; 2021.
23. Kim H. Deep video portraits. ACM Transact Graph. 2018;37(4):163.
24. Larson J, Angwin J, Mattu S, Kirchner L. Machine bias. New York: ProPublica; 2016.
25. Latif S, Qayyum A, Usama M, Qadir J, Zwitter A, Shahzad M. Caveat emptor: the risks of using big data for human development. IEEE Technol Soc Mag. 2019;38(3):82–90.
26. Liao SM. Ethics of artificial intelligence. Oxford: Oxford University Publication; 2020.
27. Lin P, Jenkins R, Abney K. Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford: Oxford University Press; 2020.
28. Martin N. Artificial intelligence is being used to diagnose disease and design new drugs. New Jersey: Forbes; 2019.
29. Université de Montréal. Montreal Declaration for responsible AI development. Montreal: Université de Montréal; 2018.
30. Mohamed, et al. Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos Technol. 2020;33(4):659–84. https://doi.org/10.1007/s13347-020-00405-8.
31. Mittelstadt B. Principles alone cannot guarantee ethical. AI Nat Mach Intell. 2019;1(11):501–7.
32. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics. 2020;26(4):2141–68. https://doi.org/10.1007/s11948-019-00165-5.
33. Newport C. When technology goes awry. Commun ACM. 2020;63(5):49–52. https://doi.org/10.1145/3391975.
34. Noble SU. Algorithms of oppression: how search engines reinforce racism. New York: New York University Press; 2018.
35. O'Neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown; 2016.
36. Pichai S. AI at Google: our principles. Mountain View: Google; 2018.
37. Postman N. Five things we need to know about technological change. Denver: University of California; 1998. p. 28.
38. Qadir J, Islam MQ, Al-Fuqaha A. Toward accountable human-centered AI: rationale and promising directions. J Inf Commun Ethics Soc. 2022. https://doi.org/10.1108/JICES-06-2021-0059.
39. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng. 2020;14:156–80.
40. Qayyum A, Usama M, Qadir J, Al-Fuqaha A. Securing connected & autonomous vehicles: challenges posed by adversarial machine learning and the way forward. IEEE Commun Surv Tutorials. 2020;22(2):998–1026.

41. Raquib A. Islamic ethics of technology: an objectives (Maqāṣid) approach. New York: The Other Press; 2015.
42. Riken. AI used to detect fetal heart problems. Helsinki: EurekAlert!; 2018.
43. Rivers TJ. Progress and technology: their interdependency. Technol Soc. 2002;24(4):503–22. https://doi.org/10.1016/s0160-791x(02)00039-8.
44. Satell G, Abdel-Magied Y. AI fairness isn't just an ethical issue. Brighton: Harvard Business Review; 2020.
45. Selbst AD. Disparate impact in big data policing. Ga L Rev. 2017;52:109.
46. Walī Allāh S. The conclusive argument from God: Shāh Walī Allāh of Delhi's Ḥujjat Allāh al-bāligha. (M. Hermansen, Trans.) (Vol. 25, Ser. Islamic Philosophy, Theology and Science. Texts and Studies). Leiden: Brill; 2020.
47. Taha A. Al-'Amal al-Dini Wa-Tajdid al-'Aql. 4th ed. Casablanca: al-Markaz al-Thaqafi al-'Arabi; 2006.
48. Taplin J. Move fast and break things: how Facebook, Google, and Amazon have cornered culture and what it means for all of us. Basingstoke: Pan Macmillan; 2017.
49. Toyama K. Geek heresy: Rescuing social change from the cult of technology. New York: PublicAffairs; 2015.
50. Tucker I. Roger McNamee: 'Facebook is a threat to whatever remains of democracy in the US.' Kings Place: The Guardian; 2020.
51. Vallor S. Technology and the virtues a philosophical guide to a future worth wanting. Oxford: Oxford University Press; 2018.
52. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, Nerini FF. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun. 2020;11(1):1–10.
53. World Bank Group. World development report 2016: digital dividends. Washington: World Bank Publications; 2016.
54. Yuan L. How cheap labor drives China's A.I. ambitions. New York: The New York Times; 2018.
55. Zicari RV, Brodersen J, Brusseau J, Düdder B, Eichhorn T, Ivanov T, Westerlund M. Z-Inspection®: a process to assess trustworthy. AI IEEE Transact Technol Soc. 2021;2(2):83–97.
56. Zuboff S. The age of surveillance capitalism: The fight for a human future at the new frontier of power. New York: PublicAffairs; 2019.