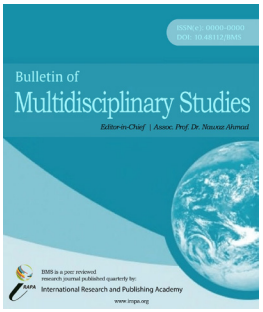
13-13-BMS-6551-1223

Bulletin of
Multidisciplinary Studies
Vol. 3 No. 1 | January – March, 2026 | 112-119
ISSN(e): 3007-0619 | ISSN(p): 3007-0600
DOI: https://doi.org/10.48112/bms.v3i1.1223



Original Article

Artificial Intelligence and Moral Agency: A Critical Philosophical Investigation

Bashir Ahmed Jatoi (Ph.D)[*1] & Ayaz Hyder Mugheri[2]

[1] *Department of History, University of Sindh, Jamshoro, Pakistan*
[2] *Department of Philosophy, University of Sindh, Jamshoro, Pakistan*

Article history:

Received: December 21, 2025
Revised: March 01, 2026
Accepted: March 03, 2026
Published: March 31, 2026

ABSTRACT

This paper carried out a philosophical analysis of the moral agency concept as applied to artificial intelligence. Based on traditional theories developed by Aristotle, Kant, and Hume, the research outlines the normative standards of agency, that is, autonomy, intentionality, and moral responsibility. It is against this theoretical scaffolding that the paper critically analyses the current AI systems and concludes that, despite the ability to simulate moral agency by using advanced rule-based systems or machine-learning algorithms, they do not possess the critical qualities of consciousness, free will, or moral self-reflection. Through the use of case studies and thought experiments, specifically the Turing Test and the Chinese Room argument, the analysis proves that the existing AI systems do not possess the ontological depth necessary to be considered as the agents of true morality. The paper is not a statement that AI is incompetent as an agency; it explores other conceptions, including distributed agency and relational agency; in particular, it explores socio-technical systems where human and machine actions are closely embedded. The ethical implications, especially in low-regulatory environments, as is the case in Pakistan, are given special consideration, where digital inequalities and pluralism across cultures can serve to expand accountability gaps. Finally, the paper will argue that making AI the subject of moral agency is not only philosophically unsound but also morally dangerous. It promotes a less carefree attitude to human responsibility and suggests extensive ethical regulation of the creation and implementation of intelligent systems.

Keywords: *Artificial Intelligence Ethics, Moral Agency, Philosophy of Technology, Postcolonial Technology Ethics, Relational Responsibility*

*Corresponding Author: Bashir Ahmed Jatoi (Ph.D) | Jatoibashir@usindh.edu.pk
JEL Classification: **O14**

How to Cite:
Jatoi, B. A., & Mugheri, A. H. (2026). Artificial Intelligence and Moral Agency: A Critical Philosophical Investigation. *Bulletin of Multidisciplinary Studies, 3*(1), 112–119. https://doi.org/10.48112/bms.v3i1.1223

Publisher's Note:
International Research and Publishing Academy (*i*RAPA) stands neutral with regard to jurisdictional claims in the published maps and institutional affiliations.

INTRODUCTION

The emergence of Artificial Intelligence (AI) as a decision-making agent in domains such as healthcare, criminal justice, warfare, and education has given rise to a new wave of ethical inquiry concerning its moral status. At the heart of these inquiries lies a fundamental philosophical question: *Can artificial intelligence possess moral agency?* This question is not merely technical or speculative; it strikes at the core of long-standing debates in moral philosophy regarding autonomy, consciousness, and responsibility. The traditional understanding of moral agency refers to the ability of an agent to make moral judgments based on a notion of what is right and wrong and to be responsible with respect to actions taken in accordance with those judgments (Burger, 2009). In order to be recognized as a moral agent, a being needs to have intentionality, autonomy and moral understanding-characteristic attributes typically attributed to rational and sentient beings. As the artificial intelligence systems grow more advanced in simulating moral judgments via an algorithmic framework and machine learning, some observers are already inclined to consider such systems autonomous moral beings (Gunkel, 2018).

However, this point of view runs a risk of confusing the artificial simulation of ethical behaviour through algorithms with a real moral thought process. Besides, the AI algorithms are vulnerable to manipulation, data leakage, and the decline in performance caused by external interference. Such vulnerabilities can pose a direct threat to human life in safety-critical areas like medicine and autonomous driving. The lack of intrinsic self-observation, security consciousness, and moral will further emphasize that AI systems run without a sense of internal moral integrity that would make them capable of having a moral agency (Huang et al., 2022).

The key question that is in this paper is to what degree can modern artificial-intelligence systems be considered to have genuine moral agency, or are their moral behaviours simply superficial simulations lacking any conscious ethical intentionality? With the classical and postcolonial theoretical frameworks in mind, the current research argues that it is a conceptually invalid and morally dangerous act to attribute moral agency to AI. The growing complexity of AI across areas that used to be viewed as distinctly human-language generation, pattern recognition, even emotional imitation has only exacerbated arguments across various fields. Within applied ethics and the philosophy of technology, researchers ask the question of whether AI is just a reflection of human moral standards or it starts to exhibit some type of autonomous moral behaviour (Coeckelbergh, 2022).

Parallel to that, legal scholars and policy makers struggle to distribute the blame in case of malfunction or harm done by the AI systems. These anxieties come to fruition outside of abstraction; they impact social ideas of fairness, responsibility and moral design. The present paper challenges the philosophical premises of the attribution of moral agency to artificial intelligence. It begins by outlining the key formulations of moral agency presented by Aristotle, Kant and Hume and thus providing a normatively solid framework with the critical evaluation of modern AI architectures. The paper then empirically operationalises this scaffold by questioning real-world examples of its implementation, such as autonomous vehicle systems, algorithmic predictive policing, and AI-enhanced healthcare, to determine whether the existent technologies meet the necessary ethical standards of agency. Based on the Chinese Room argument by Searle (1980), and classical Turing test (Turing, 2007), the author disagrees with the idea that AI can have some sort of understanding or intent in a moral setting.

The methodological paradigm applied in this study is conceptual analysis and philosophical critique and empirical examples of the Global North and the Global South are used to reduce techno-centric prejudice. Special attention is devoted to the postcolonial and Islamic ethics, especially in low-regulation societies like Pakistan, where problems of responsibility and government taking on new cultural implications (Khan & Ullah, 2024). The main argument that is put forward here is that although it might seem that artificial intelligence is autonomous in the organized settings, it does not have the ontological richness such as consciousness, freedom, and moral interiority to qualify as a moral agent. Instead of projecting moral agency onto machines, relational and distributed approaches to responsibility need to be considered, which can better explain the entanglement of human and machine behaviours in the hybrid. This method aims at maintaining moral purity but adjusting to technological sophistication of modern society.

METHODOLOGY

The research question that will be used now will take a normative approach to philosophy; it will combine both classical methods of conceptual studies with the empirically feasible case studies to determine the possibility of applying moral agency to artificial intelligence (AI). Classical conceptual analysis Classical conceptual analysis is an established approach of

analytic philosophy which considers definitions, antecedent consequences and requisite conditions of moral agency as described in the major foundational ethical theories, specifically Aristotle, Immanuel Kant, and David Hume. This question thus clarifies the key concepts of autonomy, intentionality, moral affectivity, and accountability and evaluates how much they can be extended to non-human agents, including AI systems (Hutto, 2012).

Applying the concepts, the paper incorporates reflective equilibrium where normative theory and practical cases are refuted and modified against one another to reach to a justifiable ethical position (Rawls, 1971). The case studies, including autonomous vehicles, judicial algorithms, AI in healthcare triage, and lethal autonomous weapons, are examples of empirical examples that exercise the philosophical assertions. These are not chosen on technical grounds but due to their ethical importance and to the discussion of whether machines can have the moral agency. Moreover, the paper uses an intercultural ethical approach, which relies on Islamic philosophical morality, post-colonial theory, as well as social-technology systems thinking, to challenge Eurocentric moral universalism, and to adapt to Global South contexts. This issue is especially acute in Pakistan, where the perception of AI ethics is influenced by the varying ethical ideas, as well as by the technological framework dictated by the lack of regulations (Rahman, 2001). The current research therefore challenges the post-colonial and Islamic views so that the inequalities surrounding the design, use, and management of AI technologies are reinstated.

Towards relational ethics and distributed responsibility theory, the paper also explores the distributed presence of agency and responsibility in human-machine network, institutional and design environment. Such a shift to a more relational over individualistic conceptualizations of agency is also becoming widely accepted by the philosophy of technology, due to its greater ability to reflect the complex entanglements of AI systems in the real-world context (Coeckelbergh, 2010). As a result, the methodological approach can be described best as a philosophical-ethical question, which is placed in the context of the classical moral theory but extended by critical theory and applied ethics. Its main goal is not only to oppose the ontological justification of attributing moral agency to artificial intelligence, but also to provide a more context-specific ethical approach to outlining responsibility in AI-human assemblages.

Philosophical foundations of moral agency

The idea of moral agency of Aristotle is based upon the practical reason (phronēsis) when ethical action is the result of the conscious choice supported by habitual virtue. In the Nicomachean Ethics, he considers voluntary action as a precondition to moral responsibility that only rational agents who are able to make intentional decisions based on understanding of the good could qualify to become a moral agent (Burger, 2009). Was this treated according to this Aristotelian model, artificial intelligence would have no telos or final cause, which would guide moral agents to eudaimonic flourishing. Algorithms may produce rational-appearing outputs, although these do not exist due to internalized notion of the good. Such criticism is supported by recent critiques. Johnson and Verdicchio (2017) state that artificial intelligence lacks the concept of moral character since it is based on instrumental logic, not reflective virtue of virtue (p. 578). Its action might seem complex, but it is not based on the developing sense of self or community. AI is incapable of developing virtue it merely performs the procedures created by other people.

Kantian ethics gives the strictest definition of moral agency; the capacity of acting because of a sense of duty brought by principles that are autonomously enacted by the individuals. Kant defines a moral agent in the Groundwork of the Metaphysics of Morals, as an agent who is able to will a maxim that can make a universal law (Kant, 1996). What is more important is that such agency also assumes autonomy but, in a substantive, rather than a nominal sense of being independent, that is to be governed by a rational moral law and to self-legislate. The systems of artificial intelligence, including those based on reinforcement learning or deep neural networks, lack the shape of autonomy that Kantian moral philosophy is founded on. This weakness is realized once such systems are put into practice in the field. As an example, data collection AI regularly uses personally identifiable information in order to optimize its prediction algorithms, which is a violation of the imperative of treating individuals as an end in itself in the thought of Kant. Further still, autonomous vehicular technologies that prioritize efficiency over human safety demonstrate the failure of AI to construct moral principles autonomously; instead, the described agent acts according to the external stimuli, using the pre-defined computational models.

Coeckelbergh (2022) notes that the autonomy displayed by AI is not moral but operational as it is essentially designed by humans and not by their own moral judgment. Such systems do not meet the requirements of moral agency set out by Kant in the lack of reflexive reflection on the justifications to acting thus, and of acting based on duty rather than on calculus

of consequences or utilitarian reasoning. Unlike Aristotle and Kant, David Hume believes that reason is the servant of the passions; to Hume, our morality is a product of moral sentiment that is, our emotional reactions to actions and agents. As it can be seen in the work A Treatise of Human Nature, published by Hume, morals stimulate passions, and cause or avert actions (Hume, 2000). Therefore, a moral agent should have the affective faculties, which comprise the foundation of moral judgment in the form of sympathy, empathy, and indignation.

Based on this criterion, artificial intelligence is barred twice: it cannot have emotions and understand them in the context of moral agents. In spite of the possibility of affect-recognition systems simulating the affective displays, Bryson (2018) notes that affective display is not a phenomenon of AI, but a mere simulation of the same. The lack of real emotional richness undermines the chances of a real moral motivation and consequently makes the AI behaviour ethically relevant, but not ethical. Although the modern AI shows behavioural competence the ability to cope with complicated situations, to learn, and recreate dialogue, these competencies are more of an algorithmic nature than intentional. However, several researchers disagree with these traditional demarcations. Gunkel (2018) argues that an extreme implementation of the historical ideas of moral agency can prevent the identification of new types of agencies in the areas of human-machine systems. However, he admits that modern artificial intelligence is still lacking in the area of self-understanding or reflexivity needed to attribute moral praise or blame.

Classical philosophical notions of moral agency enforce some of the strictest requirements' autonomy, rational volition and moral sentiment that modern artificial-intelligence systems are evidently not able to meet. Even though AI can perform moral-relevant actions, it is not moral since it does not apprehend or aspire in a morally relevant way. The idea of moral agency attribution to AI is, therefore, at best, a philosophical renunciation, and at worst, a dangerous shift of human responsibility. These hypotheses of theory will be subject to empirical analysis in the following sections with real-life applications of AI, and, at the same time, will pay close attention to cultural and ethical frameworks that can redefine, but not erase, the philosophical boundaries of agency. The gaps in philosophy singled out, namely: the lack of Aristotelian phronesia, Kantian autonomy and Humean moral sentiment in AI, is reflected in practical ethical levels. The theorists use salient criteria such as rational autonomy, intentionality, emotional capacity and moral accountable to demarcate moral subjects. These characteristics are not simply philosophical abstractions; they are ongoing standards of ethical agency in diverse moral systems. Despite the growing sophistication of artificial intelligence, currently existing AI systems lack ontological and normative characteristics that they need to meet these standards. The following table provides a relative synthesis of the moral agency criteria that can be derived out of the traditions mentioned above, evaluating the extent to which modern AI systems either display or lack display each of the attributes. The given visual image explains the conceptual divide between genuine moral agency and AI-mediated moral simulation and, as such, supports the argument that attributing moral agency to machines is not a constructive redefinition of agency but a category error.

Table 1
Comparative Criteria of Moral Agency in Classical Philosophy and Artificial Intelligence

Criteria	Aristotle (Virtue)	Kant (Duty)	Hume (Sentiment)	Contemporary AI
Rational Deliberation	Yes	Yes	Partial	No
Autonomy/Self-rule	Yes	Yes	Contextual	No
Moral Emotion	No	Not Central	Yes	No
Intentionality	Yes	Yes	Yes	No
Accountability	Yes	Yes	Yes	No

Artificial Intelligence in Moral Contexts

Artificial intelligence has transitioned from a theoretical domain to a practical agent of influence across sectors with profound moral implications. Its deployment in autonomous vehicles, judicial systems, healthcare, and military applications has prompted scholars and policymakers to confront an unsettling question: when AI systems make ethically significant decisions, are they reasoning morally, or merely simulating moral behaviour based on programmed logic? An example of one of the most vivid applications is autonomous vehicles, where the choice taken in a split second can result in life and death. The global Moral

Machine is an experiment that was conducted by the researchers at MIT and involved millions of participants to study how autonomous cars ought to act in the ethically difficult situations (Awad et al., 2018). The findings showed that there was vast cultural difference in moral preferences creating immediate concerns as to whether universalistic moral programming was possible. However, despite the efforts of developers to encode ethical decision rules in such systems, these vehicles are always based on predetermined risk measurements and maximization algorithms. They have no practical or deliberate understanding of the moral stakeholders involved. According to Lin (2016), what may seem moral judgment in machines, is in actuality some kind of moral calculus without conscious thought or normative comprehension.

Similar issues can be observed in the criminal justice field, where risk assessment tools powered by artificial intelligence have been implemented (e.g., COMPAS) to forecast the recidivism. Nevertheless, these tools have received criticism in the reproduction of historical biases (most of which are against racial minorities) despite their widespread adoption (Angwin et al., 2022). In contrast to the human adjudicators, AI systems do not have the ability to perceive the ideas of fairness, retribution, or rehabilitation; they just make correlations and tune predictive performance. Eubanks (2018) argues that alleged objectivity of these tools masks confronted systemic prejudices, as a result of which inequality is being mechanized practically. This fact poses some serious philosophical challenges to moral agency: without a concept of justice, is it possible to even consider that a system is acting with ethical discretion?

The problem is more acute in the area of healthcare. Artificial-intelligence systems were utilized during the COVID-19 pandemic to guide the triage decision-making process, which includes what patients are prioritized to receive care in resource-limited environments. These inherently moral decisions were performed on the basis of utility-based scoring functions. Although, these approaches enhanced the efficiency of the procedures, they also endangered human lives as just another piece of data. Cha and Kim (2025) see these choices as being sensitive to morality, especially regarding disability, age, and social-economic considerations, which AI cannot be able to sensibly record. The reliance on opaque systems of AI to make life-and-death decisions, London (2019) warns, creates a moral vacuum where neither humans nor machines can be held accountable at the end of the day and neither has a moral competence.

The issue of explicability within the field of artificial intelligence also makes the argument about moral agency attribution to such systems even weaker. Even so-called simple decision trees, which are among the most common models used in machine learning, may reach a certain level of complexity that cannot be recreated by the human user. In a demonstrative example, computer scientist Finale Doshi-Velez gave a decision tree based on only five input variables to a group of experts. Although a part of the participants was able to adhere to the logical reasoning process, most of them were unable to explain their choice in a plain and understandable language. More complex models, including deep-learning architectures of order millions of parameters, are an even more opaque. Such an experience reveals a severe lack of connectivity: where AI systems may produce outputs that seem to have a high level of precision, they do not have semantic understanding, or the ability to provide normative or moral explanations to their output. In the absence of transparency, reflective contemplation or rational substantiation, the choices that these systems render are not morally reasoned and hence the perception that the systems simulate moral conduct as opposed to modelling it.

The most ethically frightening use of artificial intelligence is in military use, especially lethal autonomous weapon systems (LAWS). They are designed in such a way that they will detect and counter their targets without direct human control. Whereas proponents also believe that this type of automation could help to minimize errors on the battlefield, opponents believe that removing human beings out of the lethal decision-making operations is a violation of moral and humanitarian values. According to Garcia (2018), it is assumed that such systems do not comply with the principles of the just war theory, in particular, such principles as proportionality and discrimination; since autonomous machines have no empathetic abilities and cannot comprehend the moral significance of killing a human being, their decisions are based on recognizing patterns and classifying targets, not morality.

Through these diverse applications there is a common denominator, AI systems do not exhibit the requisite traits of genuine moral agency. Their work is based on the code, which is created by humans, the statistical models, and optimization algorithms. As a result, they cannot be able to justify their actions, they cannot be able to consider the ethical dilemma thoughtfully, or to feel the affective conditions of guilt or empathy. According to Purves et al. (2015), to

take action due to morally correct reasons, one must have an internalized understanding of these reasons, which modern AI does not have. It is thus better that the action of such systems be described as a form of moral simulation, the simulation of ethically relevant behaviour that lacks intentionality, affective experience or the rational capacity that defines the behaviour of real moral agents. In addition, the assumption that artificial intelligence functions in an ethically or apolitically neutral field is highly mistaken. As explained by Mittelstadt et al. (2016), AI systems are infused with human values and assumptions at each stage of processes, including data selection to model training. It therefore follows that when AI appears as an independent agent, it is actually repeating value-based decisions of human beings in the guise of technical objectivity. Such mistaken identification can foster a dangerous devolution of responsibility, in which ethical responsibility is lost in a series of robes of automatization and abstraction.

Therefore, the introduction of artificial intelligence in ethically charged decision-making scenarios shows the much-needed distinction in the context of instrumental efficacy and moral deliberation. Although they have advanced computing capabilities, these systems lack conscious, wilful, and normative assessment, and hence they cannot be thought of as moral agents in any significant philosophical sense. Instead, they are used to mirror and magnify the human moral choice, and, as a result, create immediate questions about ultimate responsibility in the event of undesirable consequences. The practical implications of the moral simulation of AI may be further demonstrated in the examples of the actual fatalities of AI like in the 2018 Uber autonomous vehicle accident, which had led to the death of Elaine Herzberg. Although the car had a safety driver, artificial intelligence (controlling the autonomous system) made a deadly decision. The incident sparked ethical debate on whether it should be attributed to the programmer, the system or the human operator. These incidences expose the dangerous hypothesis that AI systems may entirely imitate moral decision-making. Instead, they are human-constructed reactions that tend to become inflexible and inaccurate when having to deal with complex or new situations. The ethical implications of AI activities are undeniably consequence based, but the agent of those activities is strictly human.

CONCLUSION

The accelerating integration of artificial intelligence into morally consequential domains, from autonomous transportation and healthcare triage to judicial sentencing and lethal warfare, has led to renewed philosophical urgency in examining the nature and boundaries of moral agency. This paper aims to investigate whether AI systems can genuinely be considered moral agents, and what is at stake, conceptually and ethically, when such a status is either affirmed or denied. With the theories of Aristotle, Kant, and Hume as our background knowledge, we can find that artificial intelligence systems do not meet the fundamental philosophical conditions of moral agency, i.e. intentionality, rational autonomy, and moral sentiment. The conception of moral action as virtue-driven deliberation directed at the good as conceived by Aristotle, the vision of self-imposed duty as envisaged by Kant and the foundation of morality in affective response as proposed by Hume all come down to the idea that morality agency requires a subjective interiority and normative comprehension that are currently missing in AI systems. Even though the AI can imitate the externally visible moral behaviours, they do so by algorithmically simulating data, without being conscious or responsible or even with moral judgment.

When looking at the real-world uses, one can discern that AI judges, autonomous vehicles, medical algorithms, and military robots operate in a moral-simulation environment, and not the moral reasoning. They maximize, predict, and follow goals, but do not justify, analyse, and understand about the ethical aspects of their efforts. The moral agency assigned to such systems is a categorical error which has serious moral and legal consequences. As illustrated, such a conflation traditionally results in moral offloading, where the software bears the responsibility and, as a consequence, a way to avoid accountability by the institutions and individuals. However, this criticism does not simply disapprove of the suggestion that artificial intelligence has moral character. Instead, it promotes paradigm shift in moral philosophy, which places relational and distributed descriptions of responsibility. Through such views, hyper-individualism of the classical agency theories is, thus, bypassed, and the emergence of moral outcomes in area like human-machine networks, organisational structure and global power formation is challenged. The relevance of this methodological position can be additionally enhanced through the prism of the postcolonial theory and the Global South, which anticipates politically, culturally, and ethically unequal conditions surrounding the entire global AI ecosystem. The blindly applied attribution of agency to machines, in the context of pluralistic ethical spaces and opaque governance frameworks like those in the case of

Pakistan, does not only indicate a misdiagnosis of the philosophical issues in question, but can also present a literal danger of reinforcing philosophical injustice and structural violence.

Finally, this research argues that the ethical values of artificial-intelligence behaviours should not shadow the more substantive ethical issues of human design, human supervision and responsibility. Despite its growing integration into the context of moral relevance, machines lack and cannot have moral agency in its traditional sense. The acknowledgment of this is not an act of technophobia but instead an ethnically based assertion that responsibility is a human prerogative because ethical action goes beyond action to include judgment, will, and the ability to engage in moral self-examination. Subsequent studies in the field of artificial intelligence ethics ought to continue to avoid reductive anthropomorphic explanations instead of delving into subtle, cross-cultural, and interdisciplinary studies. Besides, it should also place a high emphasis on developing philosophically sound and culturally specific governance frameworks. Such actions are the only way to make sure that the rise of artificial intelligence is accompanied by a decrease in moral responsibility.

Competing Interest

The authors had no competing interests.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254-264). Auerbach Publications.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15-26. https://doi.org/10.1007/s10676-018-9448-6

Burger, R. (2009). *Aristotle's Dialogue with Socrates: On the" Nicomachean Ethics"*. University of Chicago Press.

Cha, H., & Kim, J. (2025). Ethical considerations of artificial intelligence in emergency medicine for triage and resource allocation: a scoping review. *Clinical and Experimental Emergency Medicine, 12*(4), 306. https://pubmed.ncbi.nlm.nih.gov/41531409/

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235-241. https://doi.org/10.1007/s10676-010-9221-y

Coeckelbergh, M. (2022). *The political philosophy of AI: An introduction*. John Wiley & Sons.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Garcia, D. (2018). Lethal artificial intelligence and change: The future of international peace and security. *International Studies Review, 20*(2), 334-341. https://doi.org/10.1093/isr/viy029

Gunkel, D. J. (2018). *Robot Rights.* MIT Press.

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence, 4*(4), 799-819. https://doi.org/10.1109/TAI.2022.3194503

Hume, D. (2000). *A treatise of human nature*. Oxford University Press.

Hutto, D. D. (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT Press.

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines, 27*(4), 575-590. https://doi.org/10.1007/s11023-017-9417-6

Kant, I. (1996). *Groundwork of the metaphysics of morals*. Cambridge University Press.

Khan, I. A., & Ullah, A. (2024). The role of artificial intelligence in enhancing social governance: A framework for ethical implementation and policy development in Pakistan. *Journal of Management & Social Science, 1*(4), 274-289. https://doi.org/10.63075/9fzpbn74

Lin, P. (2016). Why ethics matters for autonomous cars. In *Autonomous driving: Technical, legal and social aspects* (pp. 69-85). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48847-8_4

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report, 49*(1), 15-21. https://doi.org/10.1002/hast.973

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851-872. https://doi.org/10.1007/s10677-015-9563-y

Rahman, F. (2001). *Health and medicine in the Islamic tradition: Change and identity*. ABC International Group.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424. https://doi.org/10.1017/S0140525X00005756

Turing, A. M. (2007). Computing machinery and intelligence. In *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer* (pp. 23-65). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3

Author Biographies

Bashir Ahmed Jatoi (Ph.D) is an Assistant Professor at the Department of History, University of Sindh, Jamshoro, Pakistan. He obtained his Doctoral Degree in History from Shanghai University, Shanghai, China.

Ayaz Hyder Mugheri is an Assistant Professor at the Department of Philosophy, University of Sindh, Jamshoro, Pakistan. He completed his Masters Degree in Philosophy from the University of Sindh, Jamshoro, Pakistan.