GRAND ACADEMIC PORTAL
RESEARCH JOURNALS

GAP BODHI TARU
A GLOBAL JOURNAL OF HUMANITIES
(ISSN – 2581-5857)
Impact Factor: SJIF - 5.551, IIFS - 5.125
Globally peer-reviewed and open access journal.

MORAL MACHINES: NAVIGATING THE ETHICAL CROSSROADS OF ARTIFICIAL INTELLIGENCE IN A HUMAN-CENTRED SOCIETY

Ms. Ashana Mishra

Visiting Faculty, Government Law College, Mumbai and Research Scholar, Department of Law, University of Mumbai
ashana.mishra@gmail.com

Abstract

Artificial Intelligence (AI) embeds as a quintessential fabric of the society in the present day. While this seeping of AI has been instrumental in myriad ways, the inevitable tangent has been that of the ethical dilemmas that surrounds the development and deployment of AI systems. This field of ethical dilemma demands urgent attention to avoid unprecedented consequences. AI-driven systems have been adopted across critical areas of human life, from predictive policing being used in the area of criminal justice, in finance and healthcare predominantly. While convenience and accuracy of results of such systems has been welcomed, one cannot rule out inherent aspects of human rights violations and lack of pinning accountability as major challenges. The reliance on machines to account for moral reasoning is too stretched an expectation, the question then arises is, can we judge the moral reasoning of a machine based on the biases and limitations of its creators? This paper emphasises on studying the ethical crossroads of AI where the primary intention is to assess the intersection of technological advancements with human centred values. The moral arc of AI isn't inherently embedded in it, rather it is curated by those who envision, design, regulate and routinely interact with such systems. The biggest challenge therefore lies in striking a balance between encouraging technological advancement while also upholding and safeguarding human dignity and mitigating harm in the course of innovation. AI is the future and this fact remains inevitable, the onus thus falls upon us to build responsible machines that align with human ethics. This goal has to be achieved by not merely directing the compass at technology alone but to draw a blueprint that enables those hands who shape its trajectory as well. This paper proposes an interdisciplinary approach where the idea is to harmonise the study of philosophy, law, social policy and computer science to have a prism view of the challenges and curate remedy for any consequences. The future of AI is not predetermined as its rests on our collective choice, responsibilities, actions and commitment. This paper serves as a call to action urging key players to build and shape ethical AI systems.

Keywords: *Artificial Intelligence, Moral Machines, Human-Centred AI, Ethics, Human Rights*

INTRODUCTION

Artificial Intelligence (AI) today no longer exists merely in the domain of science fiction, rather it has percolated to the fabrics of modern civilisation. AI, today is responsible in shaping our behaviour and in determining major outcomes, from covering the autonomous vehicles navigating urban landscapes to the use of predictive policing in the area of criminal law, AI has taken centre stage in all these areas. While AI has profound impact in our daily lives, yet it raises certain existential questions about morality, autonomy and human dignity. At the fore of all these concerns, lay a central question: Can an AI machine make moral decisions, and if so, who bears responsibility for those decisions? This paper is an attempt to understand the dilemmas using the doctrinal methodology by laying emphasis on legal frameworks, ethical theories and also incorporating real-world case studies. The primary objective is to question the existing assumptions and to lay a blueprint for a more human-centred framework that ensures zero compromise of fundamental ethical principles.

METHODOLOGY

This paper is premised on the doctrinal methodology that imbibes the study of legal statutes, various judicial decisions, international human rights instruments along with scholarly interpretations to basically evaluate the ethical challenges posed by AI. The aim is to critically assess how these laws and instruments in place address the ethical dilemmas that branch from the use of AI technologies. This study incorporates a blend of primary and secondary legal sources ranging from statutes and case laws to academic literature. This approach also seeks to study the philosophical frameworks like the Kantian deontology, Bentham and Mill's utilitarianism and certain virtue ethics to map out a more multidimensional analysis of AI and its moral underpinnings.

CONCEPTUAL FRAMEWORK: UNDERSTANDING AI AND ETHICS

AI is often categorised into two forms: Narrow AI and Artificial General Intelligence (AGI). Narrow AI is regarded as weak AI and is primarily designed to perform certain specific tasks such as language translation, content recommendation and image recognition. This functions at a level that has even surpassed human abilities. These metrics and the speed at which results are generated could not have been fathomed few years ago. Examples of narrow AI often incorporated in our daily lives include Netflix's algorithmic recommendations, voice assistants like Siri, Alexa and Google Translate. While weak AI systems still lack consciousness and emotions, since the sentient element has not been established yet, these systems still play a major role in impacting social behaviour since it includes handling of personal data.

In contrast to Narrow AI, there is AGI which currently remains merely as a theoretical concept. The aim in this regard is to create an AI system that would replicate human behaviour and have human-like cognitive skills. A system that would be capable of learning, reasoning and addressing various situations based on its own assessment. One that would learn from its own learnings without any human intervention. This seems daunting to envision but given the transcendental leap in our march towards advancement, might translate into a reality soon. Even if speculative for now, it raises some serious concerns about autonomy, control and exercise of moral authority.

THE ETHICAL DILEMMAS AND CHALLENGE OF MORAL INTEGRATION

As AI systems inch towards advancement, the current phase does involve the human element intervention, so it basically functions in spaces governed by human moral judgment. A pressing ethical issue here is AI autonomy and human oversight. This becomes even more pertinent when AI functions in domains with life and death consequences. To picture a self-driving car here, encountering a sudden accident scenario- perhaps synonymous with the trolley problem just in a more modern day set up, a problem that leaves us in fix as to how should we decide in a situation ensuring the best decision keeping morality in perspective. In the case of a self-driving car, how does it ensure weighing between two morally difficult outcomes, one that would save life of pedestrians if swerved in a particular direction while putting the passenger at risk, the other where the passenger has to be saved at the expense of the life of pedestrians.

This situation draws our attention to a vital consideration of the possibility of ruling out human oversight. Can that risk be taken by allowing absolute autonomy to AI driven systems? Immanuel Kant[1] always kept humans at the core of his philosophy. He believed that every human life holds inherent dignity and worth and that they should never be used a means to an end, rather humans should be regarded as ends in themselves. This thought is premised on the paradigm idea that every action should be guided by universal moral laws with rationality and respect for human autonomy at its core. This understanding, in the field of AI seeks to demand that machines should never be programmed in a way to make autonomous decisions that would be detrimental to human life. To apply this thought in the case of an autonomous vehicle, an algorithmic decision should never be one which would endanger a human life. Such programming and a precedent application of the same may cause more harm than good to human life.

Utilitarianism, on the other hand is more focussed on a consequentialist approach. Jeremy Bentham[2] and John Stuart Mill[3] argued that the moral value of any action is determined primarily by its outcomes, particularly when viewed from the perspective of maximising happiness and minimising pain. This philosophy, in the context of AI systems would demand programming of autonomous systems to make decisions ideally seeking 'greatest good of the greatest number'. In the example of a self-driving car, this ideal would demand a programming that is trained to sacrifice the life of one person if that decision results in saving the life of a large number of people. This would eventually reduce the overall harm caused by such an incident. But to completely agree with this idea may not be the right path always, especially when it comes to decisions where there may be absence of human oversight. Also, a precedent decision as this may sometimes also be counterproductive.

John Rawls'[4] emphasises on fairness as the foundational principle to any decision making. He was an ardent advocate of social systems that would benefit society's most disadvantaged members. This principle, when

[1] Immanuel Kant (born April 22, 1724, Königsberg, Prussia [now Kaliningrad, Russia]—died February 12, 1804, Königsberg) was a German philosopher whose comprehensive and systematic work in epistemology (the theory of knowledge), ethics, and aesthetics greatly influenced all subsequent philosophy, especially the various schools of Kantianism and idealism. Available at "Kant, Immanuel." *Encyclopedia Britannica*, https://www.britannica.com/biography/Immanuel-Kant. Accessed 22 Feb. 2025. "He argues that the human understanding is the source of the general laws of nature that structure all our experience; and that human reason gives itself the moral law, which is our basis for belief in God, freedom, and immortality. Therefore, scientific knowledge, morality, and religious belief are mutually consistent and secure because they all rest on the same foundation of human autonomy, which is also the final end of nature according to the teleological worldview of reflecting judgment that Kant introduces to unify the theoretical and practical parts of his philosophical system." Available at Wood, Allen W. "Immanuel Kant." *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/kant/. Accessed 22 Feb. 2025.

[2] Jeremy Bentham was an English philosopher and political radical. He is primarily known today for his moral philosophy, especially his principle of utilitarianism, which evaluates actions based upon their consequences. Available at "Jeremy Bentham." *Internet Encyclopedia of Philosophy*, https://iep.utm.edu/jeremy-bentham/. Accessed 22 Feb. 2025.

[3] John Stuart Mill (born May 20, 1806, London, England—died May 8, 1873, Avignon, France) was an English philosopher, economist, and exponent of utilitarianism. Available at "John Stuart Mill." *Encyclopedia Britannica*, https://www.britannica.com/biography/John-Stuart-Mill. Accessed 22 Feb. 2025.

[4] "John Rawls (b. 1921, d. 2002) was an American political philosopher in the liberal tradition. His theory of *justice as fairness* describes a society of free citizens holding equal basic rights and cooperating within an egalitarian economic system". Available at Freeman,

applied in the field of AI would seek ensuring that AI systems do not produce results with inherent algorithmic biases. Through human intervention, it has to be assured that AI algorithms are designed and audited in a way that any results produced would be free from discrimination. A glaring example of biasness in automation surfaced in the case[5] of Amazon.com Inc's recruitment drive where the inbuilt AI system reflected a disregard in shortlisting applications from women applicants. This was partially also because the past record of the company demonstrated metrics of male dominance in the giant tech industry. The philosophical dilemma in this domain is not merely abstract. It has real-world implications especially when AI is being incorporated in fields like criminal justice, healthcare and autonomous transportation. Traversing the ethical crossroads would thus require striking a fine balance between technological advancement and upholding core values of humanity.

WHO IS ACCOUNTABLE WHEN AI GOES WRONG?

AI systems pacing towards becoming autonomous in nature is an inevitable future. With the increasing pace, a glaring concern surfaces about accountability. This issue remains unaddressed under traditional legal frameworks that usually hold sentient beings accountable for want of intent, negligence or even direct causation. Ascribing these elements to non-sentient beings poses some difficulty and also a palpable gap. Take for example the case of the COMPAS risk assessment tool which was used in the United States to predict recidivism in the criminal justice system. An investigation led by ProPublica[6] put forth a revelation that the AI-driven system labelled Black defendants as high risk as compared to white defendants who had similar records. It would not be fair to ascribe malice to the AI system here, rather what this case alarms us about is the historical data embedded with past systemic biases which eventually produces discriminatory results.

Another major challenge pertaining to accountability is the Black Box problem- a phenomenon often involving the complexity of algorithms that makes it a task for even the developers to identify the origin of certain decisions. This algorithmic opacity becomes a greater challenge when engaged in life-altering decisions such as that of autonomous vehicles deciding on the greater harm principle, healthcare recommendations, deciding eligibility for parole or simply in approving loans based on the social credit systems.

The Apple's Siri Eavesdropping case[7] which revolved around privacy concerns was just another glaring concern about privacy concerns. In the class-action suit it was alleged that the voice assistant was intentionally recording people's conversations and that the same was being sold to advertisers for their promotional strategies as well as to influence consumers. The complaints alleged that conversational elements were reflected in the ads for things that were casually made during the conversations. This raised an alarm of the extent of data being collected and a compromise of the same. Apple denied such allegations and eventually settled the class-action lawsuit for a value of $95 million. A fair price over basic moral tangent of invading privacy? Not the best bargain may be. This case is pertinent from the standpoint of privacy. While our lives have become heavily reliant on devices operated by inbuilt AI systems, there is a looming opacity about data collection as well as compromise of our data to agencies and companies that fall short of maintaining transparency standards.

In yet another instance that tantamount to human rights violation, a face recognition tool operated by AI was used by the police department in the United States to find suspects[8]. Uploading the blurred image on a facial recognition tool produced the image of a man whom the police arrested while also keeping him behind bars for over a year. Similar instances were repeated wherein every person arrested claimed that they had nothing to do with the crime in question and that the arrests were baseless resulting in deprivation of their liberty, also a violation of the due process that forms the bedrock of basic rights. This raised a serious question on the moral compass that was seemingly missing from such technology ridden system easily overstepping the fairness ideal and pre-arrest procedural safeguards that form the most essential fabric of human rights.

THE LEGAL VACUUM: EXISTING FRAMEWORKS AND THEIR LIMITATIONS

While AI technologies have paced at an unprecedented rate, legal systems across have struggled to formulate laws and regulations that address societal, ethical, legal implications of this permeating technology. There is a palpable regulatory vacuum in this domain as most regulations focus on data concerns and fail to branch out to safeguards to challenges that stem from AI systems.

The General Data Protection Regulation (GDPR)[9] implemented by the European Union has often been hailed as the gold standard in terms of data protection and upholding privacy. The fundamental safeguards covered by

Samuel. *John Rawls*. Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Fall 2020 Edition, Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/entries/rawls/. Accessed 23 Feb. 2025.

[5] Dastin, Jeffrey. "Insight – Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters*, 11 Oct. 2018, https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/. Accessed 23 Feb. 2025.

[6] Angwin, Julia, et al. "Machine Bias: Risk Assessments in Criminal Sentencing." *ProPublica*, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 23 Feb. 2025.

[7] Stempel, Jonathan. "Apple to Pay $95 Million to Settle Siri Privacy Lawsuit." *Reuters*, 3 Jan. 2025, https://www.reuters.com/legal/apple-pay-95-million-settle-siri-privacy-lawsuit-2025-01-02/. Accessed 23 Feb. 2025.

[8] "Arrested by AI." *The Washington Post*, n.d., https://www.washingtonpost.com/podcasts/post-reports/arrested-by-ai/. Accessed 23 Feb. 2025.

[9] "General Data Protection Regulation (GDPR)." *GDPR Info*, https://gdpr-info.eu/. Accessed 23 Feb. 2025.

this regulation include ensuring transparency, allowing users to understand the trajectory of their data being collected and also to have some dominion over it. The regulation also covers 'right to be forgotten' which in the era of digital footprints is extremely crucial. It also embeds stricter consent requirements. While these safeguards hold relevance, however the regulation falls short on covering the ethical concerns that arise from the use of AI. The black box problem and its algorithmic complexity goes against the transparency assurance laid down in the regulation. The regulation also lacks provisions to address AI-accountability given the fact that automated systems have a bearing on life-altering decisions.

A pathbreaking initiative that surfaced after years of deliberation has been the European Union AI Act[10]. A one-of-a-kind maiden legislation designed specifically to address concerns arising out of the use of AI technologies. Its key features can be broadly classified into four major fields;

- Risk based classification- this covers those AI applications falling under the ambit of unacceptable risk domain that pose a clear threat to safety, livelihoods or rights of individuals. For example, the act calls for a ban on social credit systems or any AI system that perpetuate discrimination amongst natural persons. Article 5 of the Act[11] targets practices that leads to such discriminatory practices. This is an advance attempt in maintaining fairness while also eliminating systemic discrimination. However, what needs redressal is also the key issue of checking on the historically embedded systemic discrimination which reflects through the fed algorithms. The act also recognises potential harm that may be caused by high-risk AI systems especially in sensitive sectors like healthcare, law enforcement and critical infrastructure that have been made subject to the adherence to stringent compliance mechanisms.

[10] "AI Act Law." *AI Act Law*, n.d., https://ai-act-law.eu/?utm_source=gdpr-info.eu. Accessed 23 Feb. 2025.

[11] Art. 5 AI Act Prohibited AI Practices
1. The following AI practices shall be prohibited:
1. the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm;
2. the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm;
3. the placing on the market, the putting into service or the use of AI systems for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics, with the social score leading to either or both of the following:
1. detrimental or unfavourable treatment of certain natural persons or groups of persons in social contexts that are unrelated to the contexts in which the data was originally generated or collected;
2. detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity;
4. the placing on the market, the putting into service for this specific purpose, or the use of an AI system for making risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics; this prohibition shall not apply to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity;
5. the placing on the market, the putting into service for this specific purpose, or the use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage;
6. the placing on the market, the putting into service for this specific purpose, or the use of AI systems to infer emotions of a natural person in the areas of workplace and education institutions, except where the use of the AI system is intended to be put in place or into the market for medical or safety reasons;
7. the placing on the market, the putting into service for this specific purpose, or the use of biometric categorisation systems that categorise individually natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation; this prohibition does not cover any labelling or filtering of lawfully acquired biometric datasets, such as images, based on biometric data or categorizing of biometric data in the area of law enforcement;
8. the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement, unless and in so far as such use is strictly necessary for one of the following objectives:
1. the targeted search for specific victims of abduction, trafficking in human beings or sexual exploitation of human beings, as well as the search for missing persons;
2. the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or a genuine and present or genuine and foreseeable threat of a terrorist attack;
3. the localisation or identification of a person suspected of having committed a criminal offence, for the purpose of conducting a criminal investigation or prosecution or executing a criminal penalty for offences referred to in Annex II and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least four years-ibid

- Transparency compliance- Article 13[12] and Article 50[13] amongst other articles of the AI act cover the fundamentals relating to transparency. The act casts on deployers with the stringent responsibility to comply with transparency requirements and keeping the users informed on the interactions with the AI systems.
- Accountability standards- the high-risk AI systems must comply with the obligations laid down in the act that has been laid down for the key players in adhering to the obligations and in regulating impact assessments and risk management systems while also conducting regular audits.
- Human oversight- Article 14[14], a crucial safeguard incorporated is that high-risk AI systems should remain under human control and accountability. The act stipulates that AI systems be designed in a way the ensures human intervention to monitor and override automated decisions when necessary. This provision ensures that AI systems do not operate entirely in the autonomous domain where it could fundamentally impact fundamental rights, safety and freedoms of individuals.

The Act has been a breakthrough development, however the same is not free from lacunae. The complex implementation mechanism could burden businesses, there is also fear of overregulation which may handicap innovation. The definitions such as 'high-risk' remains ambiguous. The regulatory landscape also remains marred by global discrepancies since jurisdictions across differ in their way of regulating AI which potentially creates conflicts across markets.

India currently lacks a specific AI legislation but there has been no hesitance and resistance in adopting AI. In the absence of legislation, the Ministry of Electronics and Information Technology (MeitY) holds responsibility of strategizing the ethical use of AI through policies. This points at the lacunae of absence of a common platform to seek redressal in case of violations of rights by AI systems. The issue of Deepfakes is just one example that has threatened the moral contours of basic human rights.

CURATING A HUMAN-CENTRED APPROACH: PHILOSOPHICAL REFLECTIONS AND THE WAY FORWARD

A human-centred framework on AI should be rooted in moral grounding, may be even drawing inference from the Asilomar AI principles[15] which upholds human rights and prioritises societal benefit. Synonymous to these

[12] Art. 13 AI Act Transparency and provision of information to deployers

1. [1]High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately.

[13] Art. 50 AI Act Transparency obligations for providers and deployers of certain AI systems-ibid

1. [1]Providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system, unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use. [2]This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate or prosecute criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties, unless those systems are available for the public to report a criminal offence-ibid

[14] Art. 14 AI Act Human oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.
2. Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this Section.
3. The oversight measures shall be commensurate with the risks, level of autonomy and context of use of the high-risk AI system, and shall be ensured through either one or both of the following types of measures:
1. measures identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service;
2. measures identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the deployer.
4. For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate:
1. to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;
2. to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
3. to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;
4. to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;
5. to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.
5. For high-risk AI systems referred to in point 1(a) of Annex III, the measures referred to in paragraph 3 of this Article shall be such as to ensure that, in addition, no action or decision is taken by the deployer on the basis of the identification resulting from the system unless that identification has been separately verified and confirmed by at least two natural persons with the necessary competence, training and authority.
The requirement for a separate verification by at least two natural persons shall not apply to high-risk AI systems used for the purposes of law enforcement, migration, border control or asylum, where Union or national law considers the application of this requirement to be disproportionate.

[15] "Open Letter: AI Principles." *Future of Life Institute*, https://futureoflife.org/open-letter/ai-principles/. Accessed 24 Feb. 2025.

ideals is the United Nations Sustainable Development Goals that highlight the role of technology in advancing well-being of the society. A holistic framework must be premised on the value ideals as laid down by the Aristotelian philosophy of virtue ethics. Building a strong foundation on the lines of Kantian ethics and utilitarianism would resonate with the advancement and development purpose of technology keeping an individual at the core.

The rapid advancement of AI has ushered in transformative possibilities while also opening the pandoras box of ethical and moral concerns. Unchecked use of AI can raise serious concerns about violations of fundamental principles of ethical considerations. The regulatory landscape on the lines of GDPR and the EU AI Act tailor initial steps but are not free of shortcomings owing to the ethical complexities of the AI system. A holistic framework should spell out clear accountability and liability structures ensuring algorithmic transparency, fairness, respect for inherent human rights. These principles should shed guiding light at each step of AI development, starting right from development to deployment. By ensuring these steps, technology and progress would only augment human progress rather than violate or question the moral underpinnings of such advancement. The idea is to build a system that is not merely ‘intelligent’ but also ‘responsible’ and caters to the humane idea of morality and justice, as has been rightly and vehemently put forth by **Martin Luther King Jr., “The arc of the moral universe is long, but it bends toward justice.”**

WORKS CITED

[1] Asilomar AI Principles. *Future of Life Institute*, 2017, https://futureoflife.org/ai-principles/. Accessed 23 Feb. 2025.

[2] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford UP, 2014.

[3] Bryson, Joanna J. "The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation." *Technology and Regulation*, vol. 2019, no. 1, 2019, pp. 1–15, https://doi.org/10.26116/techreg.2019.001.

[4] European Commission. *Artificial Intelligence Act.* Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed 23 Feb. 2025.

[5] European Commission. *General Data Protection Regulation (GDPR).* 2018, https://gdpr.eu/. Accessed 23 Feb. 2025.

[6] Floridi, Luciano. *The Ethics of Artificial Intelligence and Robotics.* Springer, 2020.

[7] IEEE. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* IEEE Standards Association, 2019, https://ethicsinaction.ieee.org/. Accessed 23 Feb. 2025.

[8] Kant, Immanuel. *Groundwork of the Metaphysics of Morals.* Translated by Mary Gregor, Cambridge UP, 1998.

[9] Lin, Patrick, Keith Abney, and George A. Bekey. *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT Press, 2012.

[10] Mill, John Stuart. *Utilitarianism.* Hackett, 2001.

[11] ProPublica. "Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks." *ProPublica*, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 23 Feb. 2025.

[12] Rawls, John. *A Theory of Justice.* Harvard UP, 1971.

[13] Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach.* 4th ed., Pearson, 2021.

[14] United Nations. *Sustainable Development Goals (SDGs).* United Nations, 2015, https://sdgs.un.org/goals. Accessed 23 Feb. 2025.

[15] Winfield, Alan F. "Ethical Governance of AI and Robotics: A Roadmap." *Philosophical Transactions of the Royal Society A*, vol. 376, no. 2133, 2018, https://doi.org/10.1098/rsta.2018.0080.

[16] Wood, Allen W. “Immanuel Kant.” *Stanford Encyclopedia of Philosophy*, 2017, https://plato.stanford.edu/entries/kant/. Accessed 23 Feb. 2025.