Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

STUDY ON ETHICAL AND RESPONSIBLE ARTIFICIAL INTELLIGENCE

Dr. M. Kundalakesi MS(IT&M),M.Phil.,Ph.D [1] ,Priya Dharshini S[2], Sathish Kumar S[3]

Assistant professor, Department of computer Applications,

Sri Krishna Arts and Science College, Coimbatore

[2]student 3 BCA-A, Department of computer Applications, Sri Krishna arts and science college , Coimbatore

[2]student 3 BCA-A, Department of computer Applications, Sri Krishna arts and science college , Coimbatore

ABSTRACT

Artificial Intelligence (AI) has moved from a specialized technical field to a normalized component of contemporary work, leisure, and governance, driving a paradigm shift in global productivity and innovation. However, its rapid integration is fraught with significant risks, including privacy infringement, algorithmic bias, and the erosion of human agency. This report synthesizes findings from five primary research sources to define "Responsible AI" (RAI) as a socio-technical modus operandi that minimizes harm to people and the environment while aligning with human values. By examining the perspectives of AI practitioners, technical experts, and policymakers, the report identifies a complex web of distributed responsibility. It explores the technical foundations of privacy-preserving algorithms, the ethical risks specific to sensitive domains like healthcare and education, and the philosophical frameworks required to navigate a future involving increasingly autonomous systems.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

1. FOUNDATIONS AND DEFINITIONS OF RESPONSIBLE AI (RAI)

1.1 DEFINING RAI VS. ETHICAL AND TRUSTWORTHY AI:

The sources indicate that "Responsible AI" is often considered a more effective representation than "Ethical AI" or "Trustworthy AI" because it addresses the practical implementation of protections without stifling innovation. Ethical AI is often viewed as an unattainable ideal since it seeks to align machines strictly with human behavior, while Trustworthy AI relies on the subjective and potentially misleading perception of trust[1]. RAI serves as a modus operandi for the design, development, and deployment of AI systems to minimize risks to people and the environment. It focuses on the practical application of ethical, legal, and economic concerns to benefit society. By moving beyond subjective ethics, RAI ensures that organizations take active steps to safeguard individual liberties and preferences[2].

1.2 THE MATERIALIZED ACTION APPROACH:

Rooted in Science and Technology Studies (STS), this approach assumes that technological artifacts like AI give material form to social processes. AI systems are viewed as being imbued with human subjectivity, meaning the values and biases of engineers, designers, and corporate entities materialize within the artifacts. This creates a mutual constitution where people and technologies shape each other through interaction. However, the approach recognizes a subjectobject asymmetry, centralizing the responsibility of the human creator even in the face of autonomous technological forces. Politics and power are thus foregrounded in the sociotechnical system, leading to critical questions about which specific humans are held accountable.[1]

1.3 AI AS A SOCIO-TECHNICAL SYSTEM:

AI is not merely a tool but is inseparable from the socio-technical system in which it exists, involving developers, manufacturers, users, and policymakers. The technology is shaped by the social context, institutional frameworks, and cultural norms of its development[1].

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

Consequently, responsibility for the actions of an AI system is shared between the technology itself and the human organizations that organize its use. This perspective acknowledges that an AI product passes through multiple hands—from training data to end-user integration—each having formative effects. It reframes AI from a simple artifact to a complex network of actors that requires collective oversight at all stages of its lifecycle.[2]

1.4 MULTIPLE DIMENSIONS OF RESPONSIBILITY:

Responsibility in the context of AI is multifaceted and categorized into several formal types: moral, legal, and social. Moral responsibility focuses on the ethical duty to act with fairness and justice, while legal responsibility involves liability codified in civil or criminal laws. Social responsibility highlights the duty of organizations to benefit society and avoid harming communities or the environment. Professional responsibility entails adhering to industry codes of conduct and maintaining integrity. Finally, technological responsibility involves the foresight to ensure systems are safe and aligned with societal values. Together, these dimensions define the obligations of diverse stakeholders within the AI supply chain[7].

2. CORE PRINCIPLES OF RESPONSIBLE AI

2.1 TRANSPARENCY AND EXPLAINABILITY:

These principles ensure that AI decision-making processes are understandable and accessible to those affected by their outcomes. Transparency requires the disclosure of information so that users know they are interacting with an AI and understand how their data is used. Explainability focuses on the system's ability to provide clear justifications for its decisions, which helps build user trust and allows for the assessment of reliability[6]. Together, these principles enable the identification and mitigation of hidden biases and support ethical governance. They are considered essential for maintaining public trust and ensuring that AI innovations align with human values.

2.2 FAIRNESS AND ALGORITHMIC BIAS PREVENTION:

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

This principle ensures that AI systems do not discriminate based on attributes such as race, gender, or religion. AI is susceptible to biases present in training data, which can result in biased outcomes in hiring, facial recognition, and loan applications. Mitigating this requires the use of diverse datasets, rigorous testing, and continuous monitoring of system outputs. Quantitative measures, such as disparate impact and demographic parity, are used to assess the fairness of outcomes across different groups. By adhering to these standards, AI can promote equity rather than reinforcing existing societal inequalities or discriminatory practices.

2.3 PRIVACY AND DATA PROTECTION BY DESIGN:

Especially when AI relies on vast amounts of data, this principle mandates the protection of personal information and the use of secure data practices. It involves the implementation of "privacy-by-design," where data privacy is a fundamental aspect of the system rather than an optional feature. Tools such as the Data Protection Impact Assessment (DPIA) are required to evaluate applications that pose a risk to user privacy. Techniques like data anonymization, encryption, and federated learning allow AI to function while maintaining individual anonymity. This principle empowers users to have control over their data, fostering a secure and trustworthy AI ecosystem.

2.4 ROBUSTNESS AND RELIABILITY:

AI must be capable of withstanding external attacks, erroneous inputs, and unforeseen behaviours without causing harm. Robustness ensures the safety and security of the system under malicious or uncertain real-world conditions. This is critical in sensitive domains like healthcare and autonomous vehicles, where a failure could endanger human lives. Achieving reliability requires rigorous testing, quality assurance, and the quantification of uncertainty to improve transparency. Resilience mechanisms, such as backup systems, ensure that a single failure does not lead to catastrophic damage to the entire performance.

2.5 HUMAN AGENCY AND OVERSIGHT:

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

This principle emphasizes the importance of maintaining human autonomy and control over AI systems throughout their entire lifecycle. Mechanisms like Human-in-the-Loop (HITL) require direct human intervention in decision cycles, while Human-on-the-Loop (HOTL) allows for oversight during design and monitoring. These protocols ensure that AI augments human capabilities rather than replacing them, particularly in high-stakes fields like criminal justice and healthcare. Human judgment is essential for addressing complex ethical contexts that AI may lack. Reshaping this relationship helps prevent unchecked automation and ensures that technology remains under meaningful human control[7].

2.6 SOCIALLY BENEFICIAL PRACTICE:

The socially beneficial principle mandates that AI systems provide advantages to society that significantly outweigh potential risks. This is visualized as a hierarchical pyramid: functional efficiency at the base, followed by legal, ethical, and finally philanthropic responsibilities. AI should positively contribute to public health, climate change, and poverty reduction while avoiding the exacerbation of social disparities[7]. Collaboration among technologists, policymakers, and the public is necessary to ensure AI aligns with diverse societal values. By prioritizing human well-being and inclusivity, this principle aims to maintain long-term public trust and acceptance of technological advancements.

3. TECHNICAL IMPLEMENTATION AND PRIVACYPRESERVING MECHANISMS

3.1 DIFFERENTIAL PRIVACY:

This is a mathematical technique that protects individual data points by ensuring that adding or deleting a single entry has no discernible impact on research results. It is achieved by obscuring raw data or queries with randomly applied noise[3]. This allows developers to extract valuable insights from large datasets while mathematically guaranteeing the anonymity of the individuals involved. Differential privacy is instrumental in preserving user information in analytics and protecting census databases. It serves as a key algorithmic solution for harnessing the power of AI while respecting personal data protection standards.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

3.2 HOMOMORPHIC ENCRYPTION:

This transformative paradigm enables computations to be performed on data while it remains in its encrypted format. It eliminates the need to access the raw plaintext for analysis, as only the final result is decrypted to reveal significant information[3]. This is particularly valuable for secure cloud computing and privacy-preserving computations in medical research. By keeping data secure throughout the entire processing lifecycle, it marks a substantial advancement in ethical AI development. It ensures that AI can "learn" from sensitive information without the risk of exposing the original raw data points.

3.3 FEDERATED LEARNING:

This method decentralizes the data analysis process by training AI models on local devices (like smartphones) instead of transferring data to a central server. It circumvents the privacy risks associated with central data storage and the potential for intercepting data during transfer. The central AI model is updated through the learnings gathered from these devices without ever needing to exchange or store the individual raw data samples[3]. This approach is highlighted as a primary technical strategy for maintaining individual data confidentiality. It promotes a more ethical use of AI by ensuring that learning occurs across multiple devices while preserving privacy.

3.4 INTERPRETABILITY METHODS (LIME AND SHAP):

LIME (Local Interpretable Model-agnostic Explanations) simplifies complex "black box" model predictions by approximating them with simpler, understandable models. It provides a local explanation that highlights which specific features contributed most to a single prediction. SHAP (Shapley Additive explanations), based on game theory, assigns importance values to each feature to show its contribution to a specific outcome. These methods bridge the gap between sophisticated AI systems and human understanding by offering insights into internal

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

decision-making processes. They are considered essential for ensuring responsible and informed AI deployments in high-risk sectors like banking and healthcare.

4. ETHICAL RISKS AND CHALLENGES IN AI DEPLOYMENT

4.1 PRIVACY INVASION AND DATA LEAKAGE:

These risks arise from unauthorized exposure of personal, sensitive, or confidential information belonging to students, teachers, or users. Privacy violations can occur through excessive data collection, unauthorized data sharing, or mining students' emotional lives. Data leakage is often caused by insecure storage, hacking, or inadequate management by corporations and educational institutions. Such breaches compromise human sovereignty and dignity, exposing identities and reputations to harm. These issues highlight the urgent need for strict data protection legislation and informed consent at every stage of the AI lifecycle.

4.2 THE "BLACK BOX" PROBLEM:

Many AI algorithms are uninterpretable or opaque, meaning their internal decision-making logic is hidden from users and even developers. This lack of transparency makes it difficult to understand how specific conclusions are reached, which is particularly problematic in critical fields like criminal justice and healthcare. Educators and practitioners may struggle to explain the suggestions provided by AI, leading to uncertainty about the correctness or fairness of those decisions. The "black box" nature of technology can hinder effective oversight and regulatory compliance. Addressing this requires a move toward explainable AI that offers visual or straightforward insights into the algorithmic reasoning[6].

4.3 HOMOGENIZATION AND "INFORMATION COCOONS":

In education, AI can lead to "student homogenized development," where personalized algorithms push identical content to diverse learners, restricting their growth. This creates

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

"information cocoons," where users are solidified in a closed information space, only exposed to relevant content they have previously chosen or engaged with. In the long run, this blocks the transmission of heterogeneous thoughts and limits intellectual diversity. Students may gradually lose their learning autonomy and critical thinking skills as their developmental trajectories converge. This trend threatens to alienate the unique Collision of ideas essential for authentic education[4].

4.4 SYSTEMATIC DECISION-MAKING VS. INSTINCT:

One source compares the nature of AI to that of a circus tiger, arguing that just as a predator follows its nature, AI "stays true to its general programming". If an AI system harms a human while acting within its basic code, the machine itself is not to "blame," as it lacks moral agency. Unlike humans, all AI decisions are considered systematic because they are strictly governed by algorithms, even if those decisions happen rapidly. This implies that moral responsibility remains with the human creators and users who deploy the system. Ultimately, AI is a tool whose outcomes trace back to the human subjects who set its parameters[5].

4.5 THE DIGITAL DIVIDE:

The rapid integration of AI can exacerbate the digital divide, where disparities in access to devices and digital literacy widen the gap between the rich and the poor. Different regions and industries have varying degrees of digital technology application, resulting in educational, professional, and social inequalities. Major internet education platforms may exercise technological monopolies, creating data barriers that reinforce these divisions[4]. This lack of universal application can aggravate polarization, as those with access to advanced AI tools gain significant competitive advantages. Bridging this gap requires broadening access to resources and funding, particularly in remote or underserved areas[10].

4.6 ALIENATION OF HUMAN RELATIONSHIPS:

As AI takes a more prominent role, there is a risk of alienating the teacher-student relationship, where human educators are reduced to purely functional roles. The conventional ecology of

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

education—where technology is a supplemental tool—is disrupted by "human-like AI". Interactions with machines may intensify, causing interpersonal connections among humans to fade or become localized. This can result in emotional disruption, where students experience stress or anxiety and regular emotional exchanges are sidelined. Maintaining humanistic values requires a balance where AI augments rather than replaces the social presence and nonverbal intimacy of human mentors.

5. THE ROLE OF STAKEHOLDERS AND PROFESSIONAL PRACTICE

5.1 PRACTITIONERS AS MEDIATING EXPERTS:

AI practitioners act as the "connective tissue" linking the diverse actors who set parameters for technology and the end-users who engage with the products. They occupy a unique space between powerful organizations—such as clients and legislators—and the autonomous machines that they build. While they work within pre-set constraints, they reserve a portion of responsibility for themselves due to their technical expertise. Practitioners often position themselves as mediators who translate abstract mandates into tangible software and hardware. They recognize that once a product leaves the workbench, it takes on a life of its own through user interactions and machine learning.

5.2 IMBALANCES OF POWER AND KNOWLEDGE:

The relationship between AI practitioners and those who set their parameters is defined by significant imbalances of decision-making power and technical knowledge. Practitioners feel bound by the mandates and profit-driven goals of more powerful bodies, such as corporate executives or clients. At the same time, they possess a unique skillset that their clients and supervisors often do not understand[10]. This forces practitioners to act with a degree of inevitable autonomy, making technical decisions that only they can fully comprehend. This tension can lead to "loosely coordinated confusion" during the creation of a system, making the attribution of accountability complex.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

5.3 LABOR AND DEMOGRAPHIC DISPARITIES:

The field of AI is characterized by a significant demographic homogeneity, often referred to as the "white guy problem," as the sample remains predominantly male (~75%) and white (~85%). This lack of diversity plagues the AI sector and STEM fields more broadly, potentially influencing the values and biases embedded within AI artifacts. Practitioners' choices are not neutral but are imbued with their own subjective experiences and expectations. Addressing these disparities is crucial, as underrepresented participants offer a standpoint and insights unavailable to the majority. A diversified professional base is considered essential for developing AI that reflects a wider range of social values and norms.

5.4 PROFESSIONAL ETHICS VS. ORGANIZATIONAL NORMS:

AI practitioners often rely on external conventions and organizational mandates as benchmarks for their ethical decision-making. They may be guided by company mottos or philosophical frameworks like "Just War Theory" (in the case of autonomous weapons) without necessarily being able to define specific ethical frameworks themselves. While practitioners express personal moral prerogatives, their ethics are often subservient to the formal codes of compliance set by their organizations. The presumed ethics of an organization instill confidence in the practitioner, guiding them down what they believe to be ethical pathways. This reliance on organizational norms helps offload the delicate task of subjective ethical justification from the individual professional.

6. GOVERNANCE, LEGAL FRAMEWORKS, AND GLOBAL POLICY

6.1 NATIONAL POLICIES AND THE BLETCHLEY DECLARATION:

In 2024, 28 governments—including the UK, US, EU, and China—signed the Bletchley Declaration, the first international agreement to address AI risks. The declaration acknowledges that AI poses a potentially "catastrophic risk" to humanity and mandates international

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

collaboration on safety research. Governments are increasingly positioning themselves as leaders in shaping robust frameworks that protect against negative impacts while promoting innovation. This shift toward global coordination has led to the formation of the International Network of AI Safety Institutes for global cooperation. These efforts emphasize the importance of looking collectively at the risks associated with "frontier AI"—systems that could surpass human intelligence.

6.2 REGULATORY COMPLIANCE (GDPR AND AI ACT):

Frameworks like the EU AI Act represent the world's first comprehensive regulatory effort to govern AI. The General Data Protection Regulation (GDPR) imposes strict measures on AI decision-making systems, particularly regarding personal data processing and profiling. These laws require systems to collect the minimum necessary data, maintain detailed records, and ensure transparency and accountability. Regulatory mandates provide company leadership with the authority to adopt moral philosophies without needing to defend them to shareholders, as compliance becomes "the law". However, global AI services often struggle to navigate multiple, standardized data protection environments worldwide[9].

6.3 LIABILITY AND OWNERSHIP:

Current legal consensus generally holds that AI and animals cannot be copyright holders, as seen in the "monkey selfie" case where the U.S. Copyright Office ruled that only humans can create original authorship[5]. Liability in AI accidents, such as the Uber self-driving car fatality, often rests with human supervisors rather than the machine or its creators. The court in the Uber case found that because "Level 5 Full Autonomy" does not yet exist, a human must always be ready to intervene. These precedents show that autonomous decision-making does not absolve humans from responsibility; identifying the human responsible for a machine's decision remains a legal priority. Liability laws are still evolving to ensure that victims of systematic AI errors can obtain compensation.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

6.4 INTERNATIONAL STANDARDS (ISO, OECD, UNESCO):

Leading global groups, such as the OECD and UNESCO, provide expert guidelines to ensure AI systems respect human rights and democratic ideals. The OECD's revised AI principles and governance framework serve as an international benchmark for policy-making. UNESCO's "Recommendation on the Ethics of Artificial Intelligence" emphasizes the importance of human dignity, transparency, and the rule of law. Organizations are encouraged to collaborate with standard-setting bodies like the ISO to create industry-wide guidelines for data management and algorithmic fairness. These international efforts aim to promote inclusive economic growth while protecting fundamental freedoms from AI-driven harm[8].

7. DOMAIN-SPECIFIC AI ETHICS: HEALTHCARE AND EDUCATION

7.1 AI IN HEALTHCARE:

AI applications in healthcare include diagnostic support, personalized treatment planning, and mental health assistance. Because patient well-being is at stake, these systems demand rigorous adherence to responsible principles and privacy standards like HIPAA and GDPR.

Interpretability is essential, as healthcare professionals must be able to understand and validate AI-generated recommendations before integrating them with their own clinical judgment. AI algorithms must be continuously monitored for discriminatory effects on minority ethnic communities to ensure fairness and equality in care. Ultimately, a "human-in-the-loop" approach is vital, ensuring that AI augments rather than replaces clinical expertise[8].

7.2 AI IN EDUCATION (AIED):

AIED offers personalized learning and administrative efficiency but introduces risks like academic misconduct and threats to the teaching profession. Generative tools like ChatGPT pose a paradox by potentially aiding learning while simultaneously enabling plagiarism and "contract cheating". AIED can introduce unfairness through algorithmic bias, affecting grading

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

or admissions and impacting students' lives. The "black box" nature of some educational tools means educators may not understand the logic behind personalized recommendations. Responding to these risks requires enhanced AI literacy for teachers and students, as well as the development of guidelines specifically for educational settings.

7.3 EMOTIONAL DATA MINING:

The commercial mining of students' emotional lives is an emerging ethical concern, particularly when value extraction does not align with the student's best interest. AI systems used for evaluation may misjudge a learner's emotions, providing incorrect or inappropriate feedback based on facial expressions or statements. Emotional disruption can occur when AI tools cause stress or anxiety, potentially harming a student's social and psychological development. It is argue that adolescents' emotional perception can be negatively influenced by over-reliance on AI. Mitigating this requires a comprehensive approach to emotional data, ensuring that teachers prioritize humanistic care and authentic emotional connection over automated surveillance.

7.4 MITIGATING RISKS IN AIED:

Strategies for mitigating AIED risks include data life cycle control to prevent leakage and providing transparent insights into algorithmic operations. Schools are encouraged to define the scope of "reasonable AI use" in assignments and include it in academic integrity regulations. Teachers can reform student assessments by focusing on "process assessment" and non-textual formats that cannot be easily generated by AI[4]. Parents can support this by setting "AI-free" time for critical thinking and problem-solving[8]. Governments should also introduce commercial norms to limit the exploitation of educational data for profit. Collectively, these measures aim to foster a reliable and responsible AIED landscape.

7.5 RESHAPING HUMAN-AI COLLABORATION:

Successful integration of AI requires a human-centric approach that prioritizes human decisionmaking power and agency. In healthcare, professionals must retain the authority to

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

override AI recommendations. In education, developers should strike a balance between automation and preserving the autonomy of teachers and students. Human oversight helps prevent unchecked automation and ensures that technology remains aligned with ethical norms. The goal is to create a collaborative environment where AI handles data-intensive tasks, allowing humans to focus on care, emotional support, and complex pedagogical judgment. This partnership ensures that technology serves human values rather than overriding them.

8. PHILOSOPHICAL FRAMEWORKS AND FUTURE DIRECTIONS

8.1 UTILITARIANISM VS. DEONTOLOGICAL (KANTIAN) ETHICS:

Utilitarianism posits that the morally right action is the one that produces the "most good," a mathematical logic that often appeals to algorithmic designers. However, this "calculus of good" is difficult when evaluating the value of a building vs. a human life, or different groups within society. In contrast, Kantian ethics (the Categorical Imperative) argues that a moral decision must be acceptable to everyone involved, rejecting the idea of sacrificing one for the benefit of many. These conflicting philosophies result in different legal interpretations; for example, German guidelines for self-driving cars prohibit sacrificing non-involved parties regardless of the potential majority benefit.

8.2 THE TROLLEY DILEMMA IN AUTONOMOUS SYSTEMS:

The trolley dilemma is used as an ideal model to discuss the limitations of ethics in self-driving cars. In a situation where a crash is unavoidable, the AI must decide between different groups to sacrifice, such as its own passengers or pedestrians. MIT's "Moral Machine" project allows the public to analyze their own decisions in such scenarios, highlighting that each society has different opinions on the value of life. This sophisticated discussion goes beyond pure counting and requires bold, societal-level debate to define the rules for autonomous machines. Results of these discussions become "rule utilitarianism," where defined processes predict the most positive outcome.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

8.3 SYSTEM THINKING AND "OVERTRUST":

Deming's System Thinking frames an organization as an interlinked system of humans, values, and information, noting that "a bad system will beat a good person every time". Behavioral science warns of "over trust in robots," where humans follow a machine's lead even when it is clearly malfunctioning, such as during an emergency evacuation. Humans often perceive algorithms as superior in providing unbiased information or problem-solving, leading to a "perceived helplessness" where we follow recommendations without questioning them. This automation bias can lead to omission errors (failing to notice software failure) or commission errors (accepting an incorrect automated message).

8.4 ARTIFICIAL GENERAL INTELLIGENCE (AGI):

Multiple experts predict that Artificial General Intelligence (AGI)—AI that surpasses human performance on all key benchmarks—will be achieved within 5 to 10 years. This impending shift demands immediate advancements in safety and governance mechanisms for humancentric environments. The emergence of human-equivalent intelligence introduces profound philosophical questions regarding potential AI consciousness and moral status. There is also a significant concern regarding "moral hazards," where AGI systems could act autonomously in ways that contravene human interests. This timeline underscores the urgency for robust liability frameworks and safety regulations.

REFERENCES

1. Orr, W., & Davis, J. L. (2020) analyze how AI practitioners distribute ethical responsibility across sociotechnical networks in the journal Information, Communication & Society.

2. Gunasekara, L., et al. (2025) present a systematic review of the foundations and "principle proliferation" of Responsible AI in the journal Applied System Innovation.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

3. Radanliev, P., et al. (2024) discuss the implementation of ethical AI deployment and the role of privacy-preserving algorithms in Frontiers in Artificial Intelligence.

4. Zhu, H., Sun, Y., & Yang, J. (2025) examine strategies for identifying and mitigating technological, educational, and societal risks in Humanities & Social Sciences Communications.

5. Henz, P. (2021) provides a perspective on the moral and legal responsibility for systematic AI errors in the journal Discover Artificial Intelligence.

6. Ananny, M., & Crawford, K. (2018) detail the inherent limitations of using transparency as an ideal for achieving algorithmic accountability.

7. Dignum, V. (2019) defines "Responsible AI" through a human-centered approach focused on well-being and alignment with societal values.

8. UNESCO (2021) offers specific guidance for policymakers on the ethical application and pedagogical choices of AI in the education sector.

9. The European Commission (2020) proposes a white paper establishing a European approach to excellence and trust in Artificial Intelligence.

10. O'Neil, C. (2016) critiques how opaque autonomous algorithms, termed "weapons of math destruction," can encode and replicate historical power relations.