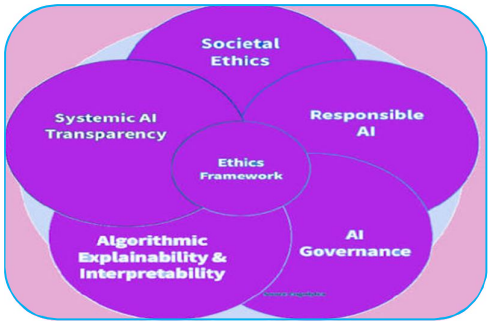
REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 15 | ISSUE - 2 | NOVEMBER - 2025

ETHICAL FRAMEWORKS FOR AI DEVELOPMENT AND INTEGRATION

Dr. Mupparaju Kishore Babu
Assistant Librarian,
Regional Institute of Education (NCERT) Mysuru.

ABSTRACT:



Artificial Intelligence (AI) has quickly become a transformative force in modern social, economic, and institutional systems. Advances in machine learning, data analytics, and autonomous decision-making have boosted efficiency and innovation across sectors such as healthcare, finance, governance, and education, but also raised ethical concerns: bias, transparency, accountability, privacy, and social justice. AI ethics now aims to ensure intelligent systems align with human values and societal well-being. This review critically examines the foundations, approaches, and relevance of the ethical frameworks that guide AI development and governance. It synthesises philosophical, technical, and policy perspectives to show how ethical principles apply in real AI systems. The study uses a systematic narrative review of journal articles, policy documents, international guidelines, and theoretical works in AI ethics, technology governance, and applied philosophy. Major sources include academic literature, regulatory frameworks, and institutional reports from global and regional bodies. The review identifies frequent themes: algorithmic bias and fairness; transparency and explainability; privacy and surveillance; accountability and liability; and broader social, cultural, and economic effects of AI. It analyses key ethical approaches: principle-based, human-centred, rights-based, risk-oriented, and governance-focused, highlighting their strengths, limits, and contexts. Findings show that no single framework is enough. A pluralistic and adaptive approach is needed for responsible AI development and governance. Embedding ethics-by-design, strengthening regulation, and fostering collaboration help ensure AI systems are trustworthy, fair, and socially beneficial.

KEYWORDS: *Ethical and Social Implications of AI; Trustworthy AI; Ethical Principles for AI; Ethics Review of AI systems.*

INTRODUCTION:

Artificial Intelligence: Changing the Way We Live and Work

Artificial Intelligence (AI) has progressed significantly since its mid-20th-century origins. Early computational models, such as Symbolic Logic (representing logic with symbols) and rule-based systems (making decisions using predefined rules) (McCarthy et al., 1955), laid the foundation for today's AI. Modern AI includes machine learning (algorithms that learn from data), neural networks

(brain-inspired information processors), and natural language processing (allowing computers to understand human language). The rise of big data (large, complex datasets) and greater computational power have accelerated AI's growth into fields like healthcare, finance, and autonomous systems (self-operating machines or software) (Bostrom, 2014). This shift has transformed industries, boosted efficiency and innovation, and raised important concerns about societal impacts. AI has improved diagnostic accuracy, optimised resource use, and changed labour markets while increasing inequalities, underscoring the need for ethical balance.

Ethical Challenges in AI Deployment

AI embedded in social decisions: AI has brought challenges as it gets integrated into social systems, leading to decisions that were previously unheard of and non-logical. One pressing issue is algorithmic discrimination, or the tendency for models trained on data with "baked in" historical biases to perpetuate past imbalances in domains such as hiring, law enforcement, and lending (O'Neil, 2016). Moreover, the "black box"-like nature of AI decision-making also undermines transparency and accountability, by making it difficult to question or audit decisions. Autonomy and surveillance concerns add another layer of complexity to the adoption of AI; facial recognition or predictive policing technologies, for instance, can have privacy and civil liberties implications (Zuboff 2019). Furthermore, automation has exacerbated social inequality due to job displacement and raised discussions on the ethical responsibility of designers and policymakers regarding such implications (Frey & Osborne, 2017).

The Necessity for Formalized Ethical Frameworks.

To address such concerns, AI development should be guided by rigorous ethical frameworks. Models should prioritise fairness, transparency, and accountability, minimising harm and maximising societal good. The EU's AI Act (EU, 2023), which requires risk assessments and human oversight for high-impact AI, exemplifies such regulation. Other efforts, such as the IEEE's Ethically Aligned Design (IEEE, 2020), call for interdisciplinary collaboration from ethicists, technologists, and communities. These approaches emphasise not only technical solutions, but also cultural and institutional changes to ensure AI reflects human values.

1.1 Scope of the Study:

This study systematically examines the ethical dimensions of Artificial Intelligence (AI) and its widespread influence on modern societies, workplaces, and decision-support systems. As AI technologies extend into sensitive domains such as healthcare, finance, governance, education, and surveillance, ethical questions involving fairness, transparency, accountability, privacy, and social justice are increasingly prominent. The study explores the philosophical underpinnings of ethics and how these moral principles are applied within AI systems. It also reviews existing ethical guidelines, regulatory models, and governance practices intended to guide responsible AI development. By integrating philosophical, technological, and policy perspectives, the study presents a multifaceted view of AI ethics and, on a global scale, details the opportunities and challenges currently shaping ethical practice.

1.2 Objectives of the Study:

1. To Identify the ethical risks of AI, including bias, discrimination, surveillance, and accountability gaps.
2. To Review current ethical concepts, governance models, and legal frameworks for AI.
3. To Discuss stakeholder roles in ethical AI and suggest future governance areas.

II. EARLIER STUDY:

Floridi et al. (2018) proposed, through the AI4People initiative, one of the main normative approaches to AI ethics. They used conceptual and comparative analysis to unite ethical principles from policy documents and various philosophies. The study identified five key principles: benevolence, non-maleficence, autonomy, justice, and explicability as the foundation of a "Good AI Society." Their recommendations stressed the need for ethics in AI to move from general principles to concrete, actionable standards. These standards should guide governance, practice, development, and institution building. The current work follows this line by discussing how AI ethics can be seen in pluralistic, principle-based terms. It also addresses real implementation challenges in different sectors. The earlier study, however, stayed mostly on a high-level, normative level, and it did not show how to apply its ideas in specific sectors. This gap requires studying how these principles work in real AI systems, which is the aim of this work.

Jobin, A., Ienca, M., & Vayena, E. (2019) conducted a landmark global mapping of AI ethics guidelines in their article. Employing a systematic content analysis of 84 ethics guidelines issued by governments, technology corporations, academic institutions, and civil society organisations from 2016 to 2019, the authors identified eleven recurring ethical principles across the global AI ethics landscape: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity. Crucially, the study found significant convergence on the top five principles — transparency, justice, non-maleficence, responsibility, and privacy — but equally significant divergence in how these principles were operationalised and prioritised across geographic and institutional contexts. The authors argued that this 'principle proliferation' without accompanying implementation guidance risks creating an 'ethics washing' phenomenon, where organisations adopt the language of ethics without substantive change to AI development practices. This study remains the most comprehensive empirical catalogue of AI ethics frameworks globally and is an essential reference for any research on ethical AI governance.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The paper made an important methodological contribution by arguing that the dominant principles-based approach to AI ethics while valuable as a normative baseline is fundamentally insufficient for guiding practical AI development decisions, because ethical principles inevitably conflict with each other in real-world applications. For example, transparency may compromise privacy; fairness to one group may produce unfair outcomes for another; autonomy may undermine safety. The authors proposed a 'tensions approach' as a complement to principles-based frameworks, requiring that AI ethics guidance explicitly map and navigate these tensions rather than treating principles as independently achievable goals. This paper has significantly influenced the methodological direction of applied AI ethics research and is foundational for researchers developing operational ethical frameworks

Hagendorff, T. (2020) published a critically evaluative study, Analysing 22 AI ethics guidelines from major technology organisations, governments, and academic bodies, Hagendorff identified significant structural weaknesses: the most technically implementable principles (such as privacy and fairness) were frequently addressed, while complex socio-political values — including democracy, collective welfare, workers' rights, and power distribution — were systematically underrepresented or absent. The study further found a stark disconnect between the ethical principles articulated in guidelines and the actual professional training, incentive structures, and corporate governance mechanisms of AI developers. Hagendorff introduced the concept of 'ethics theatre' — the performative adoption of ethical language without structural commitment — and argued for the development of binding regulatory frameworks rather than voluntary ethical guidelines. This paper is essential reading for understanding the gap between AI ethics as a normative project and AI development as a commercial and institutional practice.

Corrêa et al. (2023) conducted a comprehensive meta-analysis of 200 AI ethics guidelines and governance frameworks developed worldwide. The study identified common ethical principles including transparency, fairness, privacy, accountability, and human oversight. The authors found significant convergence among international frameworks despite differences in regulatory environments and institutional priorities. Their findings contribute to the development of globally accepted standards for responsible AI innovation. The review demonstrates that ethical AI governance requires multidisciplinary collaboration and harmonized policy approaches to address emerging technological and societal challenges effectively.

Cumming et al. (2024) explored the relationship between AI ethics and sustainability through an extensive analysis of global ethical frameworks. The study identified twenty-eight major AI ethics-led sustainability frameworks and examined their shared principles. The authors emphasized that ethical AI development should extend beyond fairness and transparency to include environmental sustainability, social inclusion, and long-term societal welfare. Their research provides valuable insights into how organizations can integrate ethical considerations into AI deployment while supporting sustainable development goals. The framework promotes responsible innovation that balances technological advancement with human and environmental well-being.

Batool, Zowghi, and Bano (2025) presented a systematic review of AI governance literature, focusing on the implementation of responsible AI principles across organizational, national, and international levels. The study analyzed governance mechanisms, stakeholder responsibilities, and ethical safeguards necessary for trustworthy AI development. The authors found that effective governance frameworks must incorporate transparency, accountability, risk assessment, and continuous monitoring throughout the AI lifecycle. Their work highlights existing gaps in governance practices and emphasizes the need for comprehensive frameworks that bridge technical, legal, and ethical dimensions of artificial intelligence management.

Papagiannidis et al. (2025) examined responsible artificial intelligence governance and proposed a comprehensive framework integrating structural, relational, and procedural dimensions of AI oversight. The study highlights the importance of transparency, accountability, stakeholder participation, and risk management in AI development. The authors argue that ethical AI cannot be achieved solely through technical solutions but requires organizational governance mechanisms that align AI systems with societal values. Their framework provides practical guidance for policymakers and organizations seeking to establish trustworthy AI ecosystems while addressing concerns related to fairness, bias mitigation, and long-term social responsibility.

III. METHODOLOGY:

This article describes itself as an analytical secondary source-based study. Journal articles, books, policy reports, ethical guidelines, international laws and regulations, as well as organizational publications concerning AI and ethics, comprise the key data sources. The approach used is thematic analysis, which serves to classify ethical issues, philosophical viewpoints, and framework-driven solutions for governing AI. The paper extends the comparative approach to assess various ethical approaches (e.g., “rules-based,” “rights-based,” and more nuanced, policy-oriented, and human-centred perspectives) across world cultures. Moreover, interpretative analysis is employed to connect areas of ethical theory with practical AI applications, providing a more comprehensive view of ethical AI development and its societal influences.

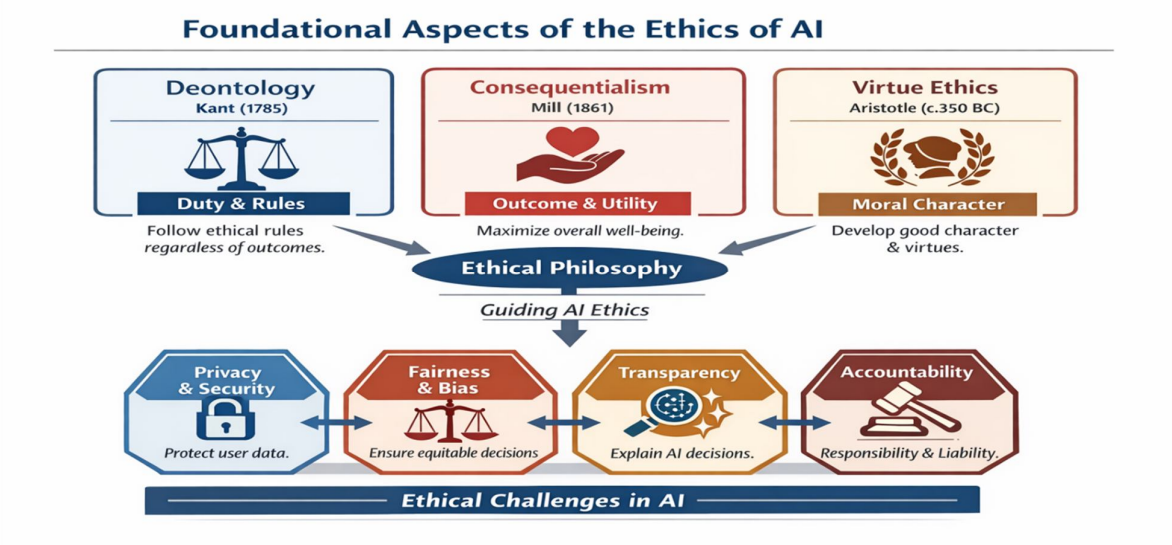


Figure.1 Conceptual Frame Work of Ethical Frameworks for AI Development

IV. FOUNDATIONAL ASPECTS OF THE ETHICS OF AI

4.1 Ethics: Definition and Philosophical Background

As the core of all subjects in philosophy, ethics researches the general principles of what is right and wrong to regulate human behaviour and decision-making. Ethics, derived from classical philosophy, can be tentatively organised around three forms of normative orientation: deontology, consequentialism, and virtue theory (Kant, 1785; Mill, 1861; Aristotle c.350 BC). Deontological formulations, such as Kant's categorical imperative, assert that rules must be followed regardless of the fruit it might yield. In particular, utilitarianism, a type of consequentialism, defines moral actions in terms of their outcomes and considers the best action to be one that maximises overall well-being. Virtue ethics, derived from The Nicomachean Ethics by Aristotle, emphasises the development of good moral habits and character. These philosophical frameworks offer critical perspectives on ethical challenges in AI. For example, deontological principles could require AI systems to inherently respect user privacy, while a consequentialist framework could evaluate the social outcomes of a decision-making algorithm. Recognising these philosophical origins is a necessary step towards the emergence of ethical rules that enable AI systems to be in tune with human values, so that cultural and moral differences are incorporated into technologies.

4.2Ethical Technology and Digital Systems

The need to incorporate ethical principles into technology and digital platforms has never been more urgent, as AI infuses many aspects of our lives, including health care, finance, and criminal justice. Technical ethics conundrums typically stem from unbalanced power, the exploitation of data, and algorithmic biases. For instance, the fact that machine learning models (often referred to as “black boxes”) are opaque raises questions of transparency, if stakeholders find it difficult to understand how decisions are reached (Binns, 2018). Historical missteps — whether it be the unintended aftermath of the Manhattan Project or misinformation spread via social media — make a case for proactive measures to safeguard ethical considerations. Models of regulation, such as the GDPR (European Commission, 2018), are being developed to make ethical considerations relevant within tech itself, with other well-known examples including the Asilomar AI Principles (Bostrom&Yudkowsky, 2014). These initiatives

focus on fairness, accountability, and transparency asking developers to balance social good with unfettered innovation. With the increasing autonomy of AI systems, the ethical obligations of technologists go beyond ensuring that their creations work; they must concern themselves with the societal and moral consequences of their decision-making tools, requiring interdisciplinary conversations among engineers, ethicists, and policymakers.

4.3 Translation from Human Ethics to Machine Ethics

Translating human morality into machine ethics is the task of capturing our ethereal moral essence in a computational model, and it is this challenge that we are wrestling with as a species. Human moral intuition is shaped by contextual variables, emotional subtlety, and cultural relativism – not easily encoded into deterministic algorithms (Wallach & Allen, 2009). For example, the decision-making of a self-driving car in an emergency situation favouring passenger safety over pedestrians, say involves ethical trade-offs that differ across philosophical and cultural traditions. There are many ways to approach machine ethics, such as rule-based systems like the Three Laws of Robotics proposed by Asimov (Naam, 2015), and more flexible methods like value alignment, where AI goals are aligned with human preferences (Russell et al., 2015). Nevertheless, issues remain to establish a universally accepted ethic and resistance to adversarial manipulation. Cooperative ventures, including AI ethics committees and participatory design processes, are beginning to emerge as means of aligning technical possibility with social values. Such a transition also requires confronting power dynamics in AI development, so that marginalised communities do not feel they are merely the objects of shaping ethical norms.

4.4 Morally Accountable AI Systems

One debate central to AI ethics is over the nature of moral agency, and thus whether AI systems could or should be considered moral agents. A morally responsible agency usually assumes purpose, knowing what one is doing, and therefore being accountable — traits that are distinctly human. Although AI systems are capable of acting accurately, their activity is non-intentional; they cannot genuinely 'morally reflect' (Floridi& Cowls, 2019). This distinction raises crucial questions of accountability: in the event that an AI system inflicts harm, who is at fault — the developer, the user, or the algorithm? The law is changing to fill these "responsibility gaps" and includes proposals for algorithmic impact assessments and liability regimes (Jobin et al., 2019). Some experts question whether granting quasi-legal personhood to advanced AI is the best way to ensure accountability, as it may oversimplify the complex interactions between human and technical biases in how AI are implemented. Finally, categorising AI as a moral patient (an entity that's morally assessed but incapable of holding moral responsibility) is ultimately an expression of the prevailing view on the topic: humans need to be more involved in governance over AIs.

V. THE IMPORTANCE OF ETHICAL FRAMEWORKS IN AI DEVELOPMENT

The rapid embedding of intelligence in crucial societal domains requires a solid ethical framework to support its development and application. That's not just an intellectual exercise that would be nice to have, but a practical necessity to manage substantial risks and ensure technology is used for the greater human good. As researchers observe, addressing these multidimensional questions is a basic requisite of responsible innovation (Floridi& Cowls, 2019).

5.1 Bias, Discrimination, and Fairness Concerns

AI models, which usually learn from historical data, tend to reflect and reinforce the biases already present in society. This can result in unfair outcomes across fields ranging from criminal justice to loan applications to hiring, making fairness a critical ethical consideration.

5.2 Privacy, Surveillance, and Data Protection

The enormous data requirements of AI pose deep privacy concerns. Modern AI systems depend on the large-scale collection, aggregation, and analysis of personal data, often gathered without the meaningful consent or awareness of the individuals concerned. This creates the conditions for pervasive surveillance, in which everyday behaviour is continuously tracked, profiled, and monetised (Zuboff, 2019). Without robust ethical guidelines and enforceable safeguards, such data can be misused for intrusive monitoring, behavioural manipulation, or discriminatory targeting. Data-protection regimes such as the General Data Protection Regulation (European Commission, 2018) seek to address these risks by establishing principles of consent, purpose limitation, data minimisation, and the right to be forgotten. Nevertheless, regulation frequently lags behind rapid technological change, and the cross-border flow of data makes consistent enforcement difficult. Strong privacy-by-design practices, transparency about data use, and meaningful user control are therefore essential to ensure that the benefits of AI do not come at the cost of individual privacy and autonomy.

5.3 Transparency and Explainability Issues

The 'black box' nature of many complex models leads to a lack of transparency, which in turn undermines trust, accountability, and the ability to detect and correct errors. The capacity of stakeholders to understand why an AI system reaches a particular decision is therefore critical to its responsible use.

5.4 Accountability and Liability Challenges

However, it is a thorny legal and ethical question of who to hold responsible when an AI fails. Frameworks would need to define where liability lies among developers, operators, and users, so there is recourse for injuries.

5.5 Social, Cultural, and Economic Implications

AI is not simply a technological problem – its social, cultural and economic effects are profound. Ethical foresight is required to prepare for possible job displacement, ensure fair access to the benefits of technology, and avoid social fragmentation.

VI. KEY ETHICS APPROACHES FOR AI DEVELOPMENT AND INTEGRATION

Historically, AI has grown from an emerging technology to a transformative force across industries globally, which in turn has raised vital ethical questions about how such systems impact our society. In response to these concerns, academics and policymakers have suggested several ethical frameworks for the responsible development and integration of AI. This article presents seven principal ethical approaches: principle-based, human-centred design and value-sensitive design, rights-based, risk-oriented and harm-minimisation, governance-oriented, and policy-oriented, explaining their key principles, practical uses, and contributions to academic debate on AI ethics.

6.1 Principle-Based Ethical Frameworks

Principle-based approaches stem from fundamental ethical ideas, including fairness, accountability, transparency, privacy, and safety. These directives serve as evaluative yardsticks to

verify that AI systems respect social values. Fairness refers to eliminating algorithmic bias so that AI does not compound or amplify bias from systems (O'Neil, 2016). Methods such as bias auditing and diverse training data sets are used to address potential forms of discrimination. Accountability also requires developers and firms to be held accountable for AI decisions, ensuring the establishment of proper lines of liability (Biddle et al, 2021). Explainability makes these decision-making processes transparent, helping ensure that decisions made by AI models are understandable to stakeholders. Privacy protects people's data, consistent with regulations such as the GDPR (European Commission, 2018). Lastly, safety and security address the avoidance of harmful side effects, ensuring AIs do what they should do without engaging in harmful behaviour or being used in unwanted ways (Russell, 2019). These values, when combined, can serve as a basis for balancing innovation with ethical duties.

6.2 Design Frameworks: Human-Centred and Value-Sensitive

HCD philosophy: Human-centred design (HCD) values inclusivity and the well-being of users by considering stakeholders' needs at all stages of AI development. Value-sensitive design (VSD), as a subclass of HCD, aims to explicitly incorporate ethical values such as autonomy, dignity, and equity into system designs (Friedman, 2020). For example, VSD could include participatory workshops with vulnerable communities to surface potential harms and benefits. This solution promotes recursive feedback loops to inform AI performance as human dynamics change. By integrating ethics early in development, VSD reduces risks and builds trust. Researchers suggest VSD is a good fit, for example, in domains such as healthcare and education, whose applications are at the intersection of human values and technical processes (Verbeek& van de Poel, 2021).

6.3 Rights-Based Ethical Frameworks

Rights-based frameworks attempt to ground AI ethics in legal and/or moral "rights," ranging from the right to privacy or freedom from discrimination to access to justice. These frameworks argue that AI should respect internationally recognised human rights, such as those contained in the Universal Declaration of Human Rights (United Nations 1948). For example, predictive policing algorithms must not infringe the 'right to a fair trial' by producing evidence-based and refutable outcomes. The proposed AI Act of the European Union (European Parliament & Council, 2023) reflects this approach, with specific regulations and compliance requirements for high-risk AI systems (e.g., law enforcement). Academics argue that rights-based models offer a common yardstick to ensure that AI follows an individual and collective rights approach (Zuboff, 2019).

6.4 Risk-Oriented and Harm-Minimisation Frameworks

Under risk-based approaches, the focus is on identifying and mitigating potential risks associated with AI use. This would involve conducting risk assessments to determine how likely and severe bad outcomes, such as job displacement, misinformation, or algorithmic errors, could be. For instance, AI systems in safety-critical domains such as healthcare or self-driving cars are subject to extensive testing for safety and the presence of fail-safes (Amodeietal, 2016). The precautionary principle is often cited, encouraging action even when scientific consensus doesn't yet exist (Bostrom&Yudkowsky, 2014). Organisations such as the OECD have produced guidelines for institutions to evaluate and mitigate the risks of AI (OECD, 2019). However, risks must not be over emphasised, as this could suppress innovation or disadvantage the developing world (Binns, 2022).

6.5 Governance and Policy-Oriented Frameworks

Governance models centre on regulatory and institutional over sight that would keep pace with an ethically driven AI. These are based on multi-stakeholder cooperation among governments,

corporations, and civil society. While the OECD AI Policy Observatory (2019) advocates transparency, inclusiveness, and international cooperation as its principles, the Global Partnership on AI (GPAI) encourages countries to discuss best practices. Policy responses, e.g., the U.S. Blueprint for an AI Bill of Rights (White House, 2022) and the EU AI Act, have also highlighted how governance bodies not only apply norms to action but also compile them into laws that shape responsible computing through legal enforcement. Good governance is the combination of top-down, stringent regulations and bottom-up standards, allowing adaptation to new technological developments (Dearden et al., 2023).

The rise of AI requires strong and effective ethical frameworks to help guide its societal influence. Principle-based frameworks set out foundational values, while human-centred and rights-based approaches aim to ensure that AI is grounded in human dignity and the requirements of the law. Risk-based regimes minimise harm, and governance structures formalise accountability. According to researchers, no single framework is sufficient; instead, a pluralistic perspective that integrates these viewpoints is required to promote AI systems that are ethical, fair and dependable (Floridi& Cowls, 2019). As AI develops, interdisciplinary cooperation and fluid regulation will also be necessary to respond to its myriad moral challenges.

VII. COMPARATIVE STUDY ON AI ETHICS OR ASSESSMENT OF OTHER FRAMEWORKS

It is important to compare ethical approaches to guide moral issues in AI. They all present different stances: utilitarianism favours solutions that maximise benefits to society (Bostrom&Yudkowsky, 2014), while deontology is concerned with following moral obligations, such as transparency and equality (Kant, 1785/1993). In contrast, virtue ethics is concerned with the development of moral character traits, such as compassion (Annas, 2011), and care ethics emphasises relational responsibilities and empathy for others in a relationship, particularly those on the periphery (Gilligan, 1982). These theoretical contrasts translated into different approaches to practice and limitations. For example, a utilitarian approach is practical and appropriate for governance-level cost-benefit analysis, but may introduce biases against minorities. Deontological rigidity works well in finance when adhering to regulations, but can be too strict. Whereas Care Ethics can relate to patient-centeredness in health AI, it cannot be scaled up in large technical systems. These domain-specific selections mirror overarching global trends; while in Western contexts, such as the OECD guidelines, they may focus on deontological values (OECD, 2019), non-Western settings tend to converge more on communal ethics (Awad et al., 2020). Through this comparative lens, the necessity of flexible, context-specific ethical approaches for AI development is revealed, as there is no one-size-fits-all framework.

VIII. ETHICAL FRAMEWORKS AT WORK IN PRACTICE: CASES FROM THE FIELD:

Ethical Theories exist as fundamental paradigms for exploring the ethical wilderness of AI. A comparison of four dominant frameworks - utilitarianism, deontology, virtue ethics, and care ethics reveals that the contributors share common ground. However, as follows: (1) the respective principles which establish these theories differ; (2) how such are considered in relation to agent and patient may diverge; (3) tensions emerge around absolutist versus relativist claims for a given principle or cluster of principles that lead theorists to operate like lone wolves with respect to their critics on various ethical topics. Utilitarianism, based on consequentialist thinking, supports AI systems that optimise total welfare (Bostrom&Yudkowsky, 2014), but its focus on aggregate outcomes may lead to the suppression of minority groups. Deontological, which is based on duty-based principles (Kant, 1993), focuses on the following rules like fairness and honesty, which are more consistent but less flexible in morally ambiguous cases. In comparison, virtue ethics focuses on moral character traits such as integrity and responsibility and is relevant to developers and users (Annas, 2011), but often provides limited prescriptive details for its use. Care ethics further departs by privileging relational obligations and

empathy, thereby becoming salient in areas with a strong focus on interpersonal relations, such as healthcare (Gilligan, 1982). Notwithstanding its strong points in relational justice, care ethics raises issues for scaling that can be particularly problematic in large technical systems. Relevance differs according to the domain: following deontological principles corresponds to regulatory requirements in finance, utilitarian cost-benefit analysis is a staple of public governance, and virtue ethics applies well to ethical literacy in education. In addition, in a generalised sense, Western constructs prevail as global norms best seen in the OECD AI principles (OECD, 2019); regional standards such as the EU GDPR embody deontological commitments, and non-Western environments are increasingly clamoring for communitarian, context-relevant perspectives (Awad et al., 2020). This comparative framing illustrates the need for pluralistic, context-sensitive ethical approaches in AI to ensure equitable, globally inclusive outcomes.

IX. ISSUES TO BE ADDRESSED WITH ETHICAL AI FRAMEWORKS

9.1 Technical Constraints

While there is agreement on the significance of ethical AI, technical limitations make it difficult to operationalize. Most AI systems are opaque, particularly the ones built on deep learning, meaning you can't audit for bias or hold anyone accountable. Explainability is still a major issue, and developers typically need to trade off model performance for interpretability (Jobin et al., 2019), making it difficult to ensure that AI systems respect ethical principles such as fairness and non-maleficence.

9.2 Organizational and Institutional Barriers

Siloed organisational structures and misaligned incentives hamper the incorporation of ethical considerations into AI development. Technical teams may not have direct access to ethics review boards, or leadership might prioritise getting to market fast over a thorough ethical review. Moreover, institutional inertia can act as a brake on the implementation of effective regulatory oversight.

9.3 Legal and Regulatory Gaps

Fast-developing AI technologies have left regulatory mechanisms in the dust. Jurisdictions are widely divergent in their approach, thus yielding an inconsistent legal terrain. Lacking harmonised regulation, organisations find it difficult to set and apply a common ethical standard, exposing them unnecessarily to the risk of non-compliance, which might cause unintentional harm.

9.4 Cultural and Contextual Diversity

Ethical standards are not universal; they differ from culture to culture and society to society. AI systems developed in one country could unintentionally violate the values of another, highlighting the importance of contextually sensitive frameworks." A solution that is one-size-fits-all may only serve to further marginalise underrepresented groups.

9.5 Ethics-Washing and Lack of Enforcement

Lastly, ethics-washing, the promotion of ethical AI initiatives without action to back it up, erodes trust. When ethical guides are voluntary, recipients of the guidance can feign compliance without consequences because such guides more often than not do not have a backup mechanism (Coeckelbergh, 2020).

9.6 Difficulties in Applying Ethical AI Frameworks

With the exception of civil liberties watchdogs like ACLU, EFF, etc., there is no clear consensus on fundamental targets for addressing problematic aspects of AI technologies.

X. STAKEHOLDERS IN ETHICAL AI DEVELOPMENT AND THE WAY FORWARD

10.1 Developers and Engineers

Developers and engineers are the building blocks of AI systems and bear considerable responsibility for incorporating ethical considerations into the design and development process. Their choices have a direct effect on the fairness, transparency and robustness of algorithms. Adhering to the principles of ethical AI, developers need to carefully identify and eliminate biases in training data, make models interpretable, and provide protections against misuse following best technical standards and continuously reflecting on the ethics of their application, they are crucial in defining trustworthy AI (Jobin et al., 2019).

10.2 Policymakers and Regulators

Policy makers and regulators play an essential role in prescribing, both legally and normatively, the contours of AI deployment. So, the European Union's proposed AI Act isn't just about regulation; it is also about the legislation of human rights and accountability, requiring transparency. Regulation supports uniform ethics across sectors, discourages potential abuse, and engenders public confidence. It is an interdisciplinary partnership that can deliver practical governance to balance innovation and societal protection, based on an understanding of how rules are to be enforced, and yet still gives them flexibility as technology evolves.

10.3 Academic and Research Institutions

Universities and other research entities, in the ethical pursuit of AI, develop theories to understand and empirically study how algorithms affect the world, as well as work in interdisciplinary isolation. They are a common space for building consensus on ethics and the long-term impacts of AI technologies. Using peer-reviewed research and curriculum development, these schools build ethical literacy in the next generation of technologists and provide a bridge between technical innovation and social responsibility.

10.4 Industry and Corporate Governance

The responsibility of global organisations and business leaders will be to infuse ethical AI principles into their business operations through robust governance models. This includes the formation of internal ethics review boards, the adoption of responsible AI charters, and supply chain accountability. Businesses need to integrate a profit motive with ethical considerations, especially in sensitive sectors like healthcare, finance and surveillance. Clear reporting and communication with stakeholders are important for maintaining public accountability.

10.5 End Users and Society

End-users or civil society have the oversight role of raising concerns, demanding transparency and accountability to developers and institutions. Citizen engagement in AI governance, whether through consultations, advocacy, or informed critique, facilitates the calibration of these systems to a range of values and the public interest. An ethically literate society has a far greater chance of safely handling some of AI's more intractable challenges, from privacy erosion to algorithmic discrimination.

XI. FUTURE DIRECTIONS AND EMERGING TRENDS

The future of responsible AI depends on several key advances. First, as generative AI systems rapidly advance, there is an urgent need for new ethical guidelines on deepfakes, misinformation and IP (Bommasani et al., 2021). Second, the incorporation of ethics-by-design models into AI development lifecycles provides a framework for proactive, rather than reactive, ethical integration. Finally, global

convergence of ethical standards implemented by international organisations (e.g., the United Nations Educational, Scientific and Cultural Organisation [UNESCO] and the Organisation for Economic Cooperation and Development [OECD]) can help minimise regulatory divergence. Fourth, we see the rise of AI auditing and certification as a means to check off ethical checkboxes. Last, but not least, an investment in educational and ethical literacy initiatives across all disciplines is crucial to providing stakeholders with the necessary skills to make decisions in AI's ethical space.

XII. CONCLUSION

Artificial Intelligence as a transformative force affecting individual and social functioning, decision-making and work organization, despite the tremendous opportunities in efficiency, creativity, and problem-solving made possible by AI-enabled technology systems, significant ethical challenges arise that should not be minimised. Concerns about algorithmic bias, opacity, and the loss of privacy, accountability, and social equity only serve to reinforce the need for clear ethical frameworks. This paper ultimately shows that there is no one recipe to tame the tangled garden of AI ethics. Rather, what is called for is a pluralistic, interdisciplinary approach that bridges philosophical ethical values and technical design architectures with legal safeguards and participatory governance. The development of AI must be accompanied by a set of principles and frameworks based on human values, enduring regulatory structures, andon-going ethical engagement. As AI technologies progress, continued cooperation amongst developers, policymakers, academics and industry is essential to ensure that AI serves the individual(s) in a fair manner consistent with society's interests.

XIII. REFERENCES

1. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
2. Brynjolfsson, E., & McAfee, A.(2017). Artificial intelligence, for real. Harvard Business Review.
3. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254–280.
4. Jobin, A., Ienca, M., & Vayena, E. (2019). The global market for AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
5. IEEE. (2020). Ethically aligned design: A vision for prioritising human well-being with autonomous and intelligent systems (2nd ed.). IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.
6. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing Group.
7. Zuboff, S. (2019). The age of surveillance capitalism. Public Affairs.
8. Aristotle (c. 350 BCE). Nicomachean Ethics, (W. D.Ross, Trans.). Oxford University Press.
9. Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), The Cambridge handbook of artificial intelligence (pp. 316–334). Cambridge University Press.
10. Binns, R. (2018). Fairness in machine learning: A political philosophy perspective. In Conference on Fairness, Accountability and Transparency, 149–159. https://doi.org/10.48550/arXiv.1806.08942
11. European Commission. (2018). General Data Protection Regulation (GDPR).
12. Floridi, L., & Cowls, J. (2019). A unified framework for the social and moral challenges of algorithms. In: Proceedings of the IEEE, 108(3), 392–404. https://doi.org/10.1109/JPROC.2019.2957445
13. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

14. Kant, I. (1785). Foundations of the Metaphysics of Morals (M. J. Gregor, Trans.). Cambridge University Press.
15. Mill, J. S. (1863). Utilitarianism. Parker, Son, and Bourn.
16. Wallach, W., & Allen, C. (2009). Moral machines: How to teach robots right from wrong. Oxford University Press.
17. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1
18. Amodei,D., Olah, C., Steinhardt, J., Christiano, P., Mané, D., and Shalev-Shwartz, S. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
19. Biddle, J., Brundage, M., Avin, S., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2021). Alignment for advanced AI. arXiv preprint arXiv:2106.11706.
20. Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), The Cambridge handbook of artificial intelligence (pp. 316–334). Cambridge University Press.
21. European Commission. (2018). General Data Protection Regulation (GDPR). https://ec.europa.eu/info/law/law-topic/data-protection_en
22. Friedman, B. (2020). Human-centred AI and value-sensitive design. AI & Society, 35(3), 579–592. https://doi.org/10.1007/s00146-020-00990-1
23. Floridi, L., and Cowls, J. (2019). An integrated set of five guiding principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.e3fdcac6
24. O'Neil, C. (2016). Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Publishing Group.
25. OECD. (2019). OECD principles on artificial intelligence. https://one.oecd.org/document/DAF/COMP(2019)13/fr/pdf
26. Russell, S. (2019). Human-compatible AI and the problem of control. Vintage.
27. United Nations. (1948).Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights
28. Verbeek, P. P., & van de Poel, I.(2021). Value-Sensitive Design as a concept in the philosophy of technology. AI & Society, 36(1), 1–11. https://doi.org/10.1007/s00146-020-00998-y
29. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Public Affairs.
30. Annas, J. (2011). Virtue ethics. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-virtue/
31. Bostrom,N., &Yudkowsky, E. (2014). The ethics of artificial intelligence. (Eds.),TheCambridge Handbook of Artificial Intelligence (pp. 316–334). Cambridge University Press.
32. Gilligan, C. (1982). In a different voice. Harvard University Press.
33. OECD. (2019). OECD AI principles.
34. https://one.oecd.org/document/DAI/NOTE(2019)11/en/pdf/
35. Coeckelbergh, M. (2020). AI ethics. MIT Press.
36. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.
37. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People's ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707.
38. Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.
39. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

40. Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 105–114.
41. Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019), 195–200.
42. Papagiannidis, E., Mikalef, P., Conboy, K., & Wilson, D. (2025). *Responsible artificial intelligence governance: A review and research framework*. Journal of Strategic Information Systems, 34(1), 101885. https://doi.org/10.1016/j.jsis.2024.101885
43. Corrêa, N. K., de Oliveira, M. A., Cohn, G., & Santos, F. (2023). *Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance*. Patterns, 4(10), 100857. https://doi.org/10.1016/j.patter.2023.100857
44. Cumming, D., Grover, P., Johan, S., & Pant, A. (2024). *Towards AI ethics-led sustainability frameworks and toolkits*. Smart Sustainable Systems, 1(1), 100003. https://doi.org/10.1016/j.ssus.2024.100003
45. Batool, A., Zowghi, D., & Bano, M. (2025). *AI governance: A systematic literature review*. AI and Ethics, 5(1), 1–25. https://doi.org/10.1007/s43681-024-00653-w