Ethical AI Decision Making Framework

PAVITRA SHARMA
AIT – CSE (Artificial Intelligence and Machine Learning)
CHANDIGARH UNIVERSITY
PUNJAB, INDIA
sharmapavitra316@gmail.com

TUSHAR SAINI
AIT – CSE (Artificial Intelligence and Machine Learning)
CHANDIGARH UNIVERSITY
PUNJAB, INDIA
tusharsainii8368@gmail.com

ARYAN CHAUHAN
AIT – CSE (Artificial Intelligence and Machine Learning)
CHANDIGARH UNIVERSITY
PUNJAB, INDIA
aryanchauhan20042003@gmail.com

Abstract—The use of Artificial Intelligence (AI) in Digital technologies (DT) is proliferating a profound socio-technical transformation. Governments and AI scholarship have endorsed key AI principles but lack direction at the implementation level. Through a systematic literature review of 19 papers, this paper contributes to the critical debate on the ethical use of AI in DTs beyond high-level AI principles. To our knowledge, this is the first paper that identifies 14 digital ethics implications for the use of AI in seven DT archetypes using a novel ontological framework (physical, cognitive, information, and governance). The paper presents key findings of the review and a conceptual model with twelve propositions highlighting the impact of digital ethics implications on societal impact, as moderated by DT archetypes and mediated by organizational impact. The implications of intelligibility, accountability, fairness, and autonomy (under the cognitive domain), and privacy (under the information domain) are the most widely discussed in our sample. Furthermore, ethical implications related to the governance domain are shown to be generally applicable for most DT archetypes. Implications under the physical domain are less prominent when it comes to AI diffusion with one exception (safety). The key findings and resulting conceptual model have academic and professional implications.

I. INTRODUCTION

The rapid integration of Artificial Intelligence (AI) into digital technologies raises significant ethical concerns at the intersection of technological progress and societal well- being. The historical lack of diversity among AI developers, coupled with the potential for fully autonomous systems to induce human dependency, poses a critical challenge. Instances of biased AI design diverting from its intended purpose and recent ethical breaches underscore the urgency of addressing these issues. The central problem lies in understanding and mitigating the ethical implications associated with AI in digital technologies, necessitating a comprehensive exploration and synthesis of existing literature to inform responsible AI integration strategies.

The increasing integration of Artificial Intelligence (AI) into digital technologies raises significant ethical concerns, as highlighted by historical diversity issues in AI development and the potential for human dependency on autonomous systems. Instances of biased AI design diverting from its intended purpose underscore the need to address these challenges urgently. The core problem revolves around understanding and mitigating the ethical implications associated with AI in digital technologies. To tackle this, a comprehensive exploration of existing literature is essential, aiming to inform strategies for responsible AI integration and development within the digital landscape.

Problems with AI

1) Risk of AI following Ethics

Artificial intelligence decision making is based on limited data, programs, relevant algorithms, and other input conditions to develop the best possible strategy. However, technology itself comes with uncertainty, and coupled with the incomplete nature of the data, decisions that lack human emotions within them are subject to decision errors and may also largely alter even human decisions, resulting in ethical risks such as privacy breaches, risk to human life, and undermining social justice; these uncertainties are an important source of ethical risks.

2) An Overview of issues

The sources of ethical risks in AI decision making include two major causes of risk: technological uncertainty and human limited rationality.

Risks in Ethical AI according to our model:

1. Fairness: Unequal power relations. Misuse of personal data. Negative impact on justice system.

2. Privacy: AI technologies often collect and analyse large amounts of personal data, raising issues related to data privacy and security. To mitigate privacy risks, we must advocate for strict data protection regulations and safe data handling practices.

3. Security: As AI technologies become increasingly sophisticated, the security risks associated with their use and the potential for misuse also increase. Hackers and malicious actors can harness the power of AI to develop more advanced cyberattacks, bypass security measures, and exploit vulnerabilities in systems.

4. Bias and Discrimination: AI systems can inadvertently perpetuate or amplify societal biases due to biased training data or algorithmic design. To minimize discrimination and ensure fairness, it is crucial to

invest in the development of unbiased algorithms and diverse training data sets.

5. Transparency: Lack of transparency in AI systems, particularly in deep learning models that can be complex and difficult to interpret, is a pressing issue. This opaqueness obscures the decision-making processes and underlying logic of these technologies. When people can't comprehend how an AI system arrives at its conclusions, it can lead to distrust and resistance to adopting these technologies.

6. Legal Challenges: It's crucial to develop new legal frameworks and regulations to address the unique issues arising from AI technologies, including liability and intellectual property rights. Legal systems must evolve to keep pace with technological advancements and protect the rights of everyone.

7. Misinformation and Manipulation: AI-generated content, such as deepfakes, contributes to the spread of false information and the manipulation of public opinion. Efforts to detect and combat AI-generated misinformation are critical in preserving the integrity of information in the digital age.

II. Defining Ehical AI

1) What Are Ethics?

AI ethics are the set of guiding principles that stakeholders (from engineers to government officials) use to ensure artificial intelligence technology is developed and used responsibly. This means taking a safe, secure, humane, and environmentally friendly approach to AI.

A strong AI code of ethics can include avoiding bias, ensuring privacy of users and their data, and mitigating environmental risks. Codes of ethics in companies and government-led regulatory frameworks are two main ways that AI ethics can be implemented. By covering global and national ethical AI issues, and laying the policy groundwork for ethical AI in companies, both approaches help regulate AI technology.

2) Existing System

The existing system on Ethical AI Framework is shaped by various guidelines and principles from reputable sources in the technology and ethics domains. Organizations such as ACM, IEEE, and the OECD have developed ethical codes, emphasizing responsible AI development. Frameworks like "Ethically Aligned Design" and the "Montreal Declaration for Responsible AI" provide comprehensive principles for creating a positive AI society. The European Commission's "Ethical Guidelines for Trustworthy AI" and the Beijing AI Principles contribute to the global landscape of AI ethics. Researchers have also examined the ethical implications of AI, with works such as "Ai4People" offering a framework for evaluating opportunities, risks, and ethical considerations. Integration of responsible practices, anticipation of impacts, and a focus on human wellbeing are key elements in these ethical AI frameworks, reflecting the evolving discourse on ethical considerations in AI development.

3) Proposed System

The proposed system for Ethical AI Framework introduces a comprehensive development process that integrates ethical analysis and prioritizes human wellbeing at each stage. Building upon existing design processes, particularly those emphasizing anticipation, reflexivity, inclusivity, and responsiveness, the framework augments these with dedicated phases for ethical impact analysis and wellbeing support. The responsible design process spans research, ideation, prototyping, and testing, with a post-launch evaluation phase added. Wellbeing, especially psychological wellbeing, is a focal point, utilizing evidence-based methods grounded in psychology. The framework acknowledges the necessity of ethical analysis, drawing from philosophy and other disciplines, to address trade-offs and values-based decisions. By embedding deliberation within the design and innovation process, the proposed system aims to ensure responsible and ethical AI development from inception through deployment.

III. Framework

Principles: While there are many existing policies and frameworks to take inspiration from, we will need to articulate our own set of principles. These principles should be aligned with the approach we take to data, including, importantly, the CEB-approved Principles on Personal Data Protection and Privacy and Data Governance policies and guidelines.

Assessment Method: The Z-Inspection process is a well-thought-out method, based on the AI HLEG principles, and the Canadian government has created an "Algorithmic Impact Assessment (AIA)" — an online questionnaire with 60 questions related to your business process, data, and system-designed decisions. Assessments can be done both at the start of a project (example: a risk-assessment that will help decide whether the project should go ahead, and which risks need to be mitigated), before an AI is deployed or periodically on an AI that is in operation — to ensure they continue to meet the guidelines. For great future coverage one should also pay close attention to the AI The Artificial Intelligence Act responds directly to citizens' proposals from the Conference on the Future of Europe (COFE), most concretely to proposal 12(10) on enhancing EU's competitiveness in strategic sectors, proposal 33 on a safe and trustworthy society, including countering disinformation and ensuring humans are ultimately in control, proposal 35 on promoting digital innovation, while ensuring human oversight and trustworthy and responsible use of AI, setting safeguards and ensuring transparency, and

proposal 37 on using AI and digital tools to improve citizens' access to information, including persons with disabilities.

Architectural Standards: We need minimum standards that specify the technical requirements AI systems, whether developed in-house or acquired as product or service, must comply with. Certain principles such as robustness or safety and security would lend themselves particularly to close governance from an architectural perspective. The architectural standards would take a role in procurement as well. The WEF "AI Procurement in a box" toolkit includes ways to integrate assessments into a procurement process.

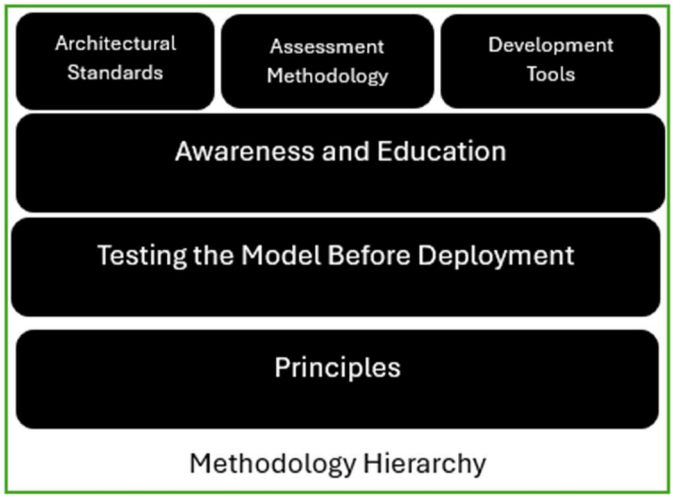
Testing of the AI Model before deployment: This particular point cannot be more emphasized on more while creating the paper as we can now also see that the Laws that are being passed by the government of different countries asking the developers to testing the AI model thoroughly before deployment of the same, as the famous english proverb goes "Prevention is Better than Cure", which makes the point of testing before deploying even greater.

Development Tools and Methodologies: These are meant to enable developers, integrators or users of AI systems to implement the principles. For instance, methodologies such as Ethics-by-Design and Value-Sensitive Design (VSD) can help to create AI in a way that leads to more accountability, responsibility and transparency. A recent paper proposes a modified form of VSD to use the Sustainable Development Goals (SDGs) as a base for ethical principles for AI that not only prevent harm but aim to do good. In terms of tools, several vendors of AI services and platforms have realized the demand for ethically aligned development of AI and are offering specific tools. Many open-source tools also exist, such as LIME, a popular algorithm for AI explainability.

Awareness and education: This should include different levels. Anyone interacting with AI systems should be provided with information in non-technical language that explains the principles and our approach. Developers, data processors and certain users of the AI systems will need more detailed documentation on our principles including technical specifications as well as general education in Ethics and training in specific tools or methodologies. Assessing the implications of an AI system can be far from trivial and specific education on this will be needed.

Governance: Our Ethical AI principles should result in a policy in the entity where it is implemented. Enforcement of this policy could happen as part of a project approval process, an architecture review board or it could be linked to data governance. The technically complex nature of AI systems will often require a balance between oversight (by auditors, senior management or other bodies) and the engineers developing or installing the AI. General-purpose AI (GPAI) systems, and the GPAI models they are based on, must meet certain transparency requirements, including compliance with EU copyright law and publishing detailed summaries of the content used for training. The more powerful GPAI models that could pose systemic risks will face additional requirements, including performing model evaluations, assessing and mitigating systemic risks, and reporting on incidents.

Transparency requirements: Additionally, artificial or manipulated images, audio or video content ("deepfakes") need to be clearly labelled as such. [8]



IV. Effectiveness

Can we be certain that these endeavors by developers and users of AI systems to implement the principles genuinely result in AI that is safe, reliable, fair, or aligns with our objectives? Many companies have faced criticism for merely paying lip service to AI ethics ("ethics-washing") while making minimal changes beneath the surface. However, these companies must balance ethical considerations with commercial interests. Fortunately, we do not encounter this tension.

Nonetheless, despite sincere efforts, the desired outcomes may not materialize without careful consideration. It is crucial not only to establish policies, standards, methodologies, and tools but also to cultivate a mindset among staff that comprehends and is dedicated to the essence of the principles. In this regard, communication and education must play a significant role in our progression.

V. Conclusion

AI technology has advanced significantly in recent years and is expected to continue evolving. With our lives becoming increasingly digitized, AI is poised to become a ubiquitous tool in various devices and services used daily.

The utilization of AI is expanding within various sectors, and its complexity and scope are likely to increase in the future. However, there are numerous ethical concerns associated with its development and use, prompting the need for careful consideration on how we integrate it into our lives. Establishing a framework of principles for ethical AI is essential to mitigate potential negative impacts.

Identify applicable funding agency here. If none, delete this text box.

These principles must align with the ethical values of operators, owners, users, and other stakeholders. Existing frameworks, such as the UN Charter and the Universal Declaration of Human Rights, provide a basis for these principles, alongside policies and frameworks developed by governments, regional organizations, and research groups.

Given AI's close relationship with data, ethical principles for AI must harmonize with those governing data, including principles on personal data protection and privacy.

Implementing ethical AI principles requires the establishment of policies, standards, methodologies, and tools. However, awareness and education are paramount. Success in implementing ethical AI principles hinges on ensuring that their essence is comprehended and embraced by all individuals interacting with AI systems.

A. Paper Reviewed

Year of Citation	Article / Author	Technique	Source
2023	Fairness in Design: A Framework for facilitating Ethical Artificial Intelligence Designs/ Jiehuang Zhang, Ying Shu, Han Yu	Fairness in design	Reference [1]
2021	Ethical Management of Artificial Intelligence/ Alfred Benedikt Brendel, Milad Mirbabaie, Tim-Benjamin Lembcke, Lennart Hofeditz	EMMA Framework (Ethical Management of Artificial Intelligence)	Reference [2]
2018	AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations/Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer Peggy Valcke, Efy Vayena	Laying Foundation for Good AI Society	Reference [3]
2023	Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI/Michael Mylrea, and Nikki Robinson	Conant's Entropy Lens	Reference [4]
2019	Towards a Framework for Ethical Audits of AI Algorithms/Ryan C. LaBrie, Gerhard H. Steinke	PAPA Framework (Privacy, Accuracy, Property, Accessibility)	Reference [5]
2019	Ethical Framework for Designing Autonomous Intelligent Systems/ Jaana Leikas, Raija Koivisto and Nadezhda Gotcheva	Ethical Considerations in Artificial Intelligence and Autonomous Systems	Reference [6]
2022	Ethical framework for artificial intelligence and digital technologies/ Ashok, Mona ORCID, Madan, Rohit, Joha, Anton and Sivarajah, Uthayasankar	Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) Framework	Reference [7]

ACKNOWLEDGMENT

We extend our sincere gratitude to Ms. Aarti for her invaluable guidance and support throughout the development of our project on **"Ethical AI Decision Making Framework"**. Her expertise and insightful feedback have been instrumental in shaping our approach and ensuring the success of our endeavor. We are truly grateful for her unwavering dedication and encouragement.

R. B. G. thanks Ms. Aarti for her invaluable contributions to our project.

REFERENCES

[1] Fairness in Design: A Framework for facilitating Ethical Artificial Intelligence Designs/ Jiehuang Zhang, Ying Shu, Han Yu , https://ieeexplore.ieee.org/abstract/document/10091496

[2] Ethical Management of Artificial Intelligence/ Alfred Benedikt Brendel, Milad Mirbabaie, Tim-Benjamin Lembcke, Lennart Hofeditz, https://www.mdpi.com/2071-1050/13/4/1974

[3] AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations/Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo,

Francesca Rossi, Burkhard Schafer Peggy Valcke, Efy Vayena, https://link.springer.com/chapter/10.1007/978-3-030-81907-1_3

[4] Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI/Michael Mylrea, and Nikki Robinson, https://www.mdpi.com/1099-4300/25/10/1429

[5] Towards a Framework for Ethical Audits of AI Algorithms/Ryan C. LaBrie,GerhardH.Steinke,https://scholar.archive.org/work/wv7pfbs3uve5xegh336ffeeuta/access/wayback/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1398&context=amcis2019

[6] Ethical Framework for Designing Autonomous Intelligent Systems/ Jaana Leikas, Raija Koivisto and Nadezhda Gotcheva, https://www.mdpi.com/2199-8531/5/1/18

[7] Ethical framework for artificial intelligence and digital technologies/ Ashok, Mona ORCID, Madan, Rohit, Joha, Anton and Sivarajah, Uthayasankar, https://www.sciencedirect.com/science/article/pii/S0268401221001262

[8] Artificial Intelligence Act: MEPs adopt landmark law https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law#:~:text=The%20Artificial%20Intelligence%20Act%20responds,including%20countering%20disinformation%20and%20ensuring