Evolving AI Ethics Frameworks and Regulations: Navigating Privacy, Safety, and Compliance in the Era of Advanced AI Tools

Roshan T Baby, Virenn Vaatts Jay

SCSEA, DYPIU, Akurdi,

DOI: 10.29322/IJSRP.14.10.2024.p15446

Paper Received Date: 21st September 2024
Paper Acceptance Date: 24th October 2024
Paper Publication Date: 30th October 2024

Abstract- This literature review explores the current landscape of AI ethics frameworks, regulations, and privacy policies to understand their roles in shaping responsible AI development and deployment. It critically examines various AI ethics frameworks, highlighting their advantages, limitations, and the challenges faced in their adoption. The review also delves into the evolving regulatory landscape surrounding AI, analyzing key regulations and policies implemented by governments and institutions worldwide to ensure ethical AI practices. Furthermore, it investigates the privacy and safety measures adopted by leading technology companies, focusing on their efforts to comply with ethical standards and safeguard user data. By synthesizing these perspectives, the review aims to provide a comprehensive overview of the progress, gaps, and future directions in the pursuit of ethically aligned AI technologies, emphasizing the need for robust frameworks and regulations to address emerging ethical dilemmas. It also builds on the work of Prem "From Ethical AI Frameworks to Tools: A Review of Approaches" by including frameworks and guidelines that were introduced after the publication of that paper.

Index Terms- About four keywords or phrases in alphabetical order, separated by commas. Keywords are used to retrieve documents in an information system such as an online journal or a search engine. (Mention 4-5 keywords)

I. Introduction

With the rapid evolution in the capabilities of AI systems, the magnitude of its impact on socio-economic, political, healthcare domains and employment, whether detrimental or beneficial, has skyrocketed. Consequently, the spotlight on its ethics has never been brighter resulting in over 200 frameworks, policies, laws, and acts.

The rise of generative AI models like OpenAI's GPT, alongside advancements from other AI developers, has not only pushed the boundaries of what these systems can achieve but also made sophisticated AI more accessible to the general public. The democratization of AI technology has fueled its integration into countless applications, ranging from automated customer service to creative content generation. However, this ease of access has amplified concerns about the misuse of AI technologies and the lack of adequate regulatory frameworks to mitigate associated risks. These concerns have underscored the urgent need for robust regulations that ensure AI is developed and deployed ethically.

Over the past few years, numerous efforts have been made to address these concerns, with more than 200 frameworks, policies, laws, and acts related to AI ethics and regulation emerging across the globe. Prior research has offered various perspectives on AI regulation, including ethical guidelines, legal frameworks, and recommendations for best practices. Notable contributions in this area include works by scholars like Bostrom (2014) on the existential risks posed by AI, Hagendorff's (2019) analysis of AI ethics guidelines, and Calo's (2017) assessment of regulatory gaps. The objective of this literature review is to build on the work of these previous studies by conducting a systematic analysis of AI ethics frameworks, with a focus on how they have evolved in response to the rapid advancements in AI capabilities. We aim to review the quantifiable impact that existing frameworks and regulations have had on the design and implementation processes of AI technologies. Additionally, this review will provide an updated list of the latest AI regulations, highlighting their key actionable points and takeaways, and will categorize and evaluate these regulations based on established criteria and our own analysis.

Ultimately, our goal is to create a comprehensive resource that captures the current state of AI ethics and regulation, identifies gaps in the existing frameworks, and outlines potential future directions for the development of ethically aligned AI technologies. By doing so, we hope to contribute to the ongoing discourse on AI ethics and emphasize the importance of establishing robust frameworks and regulatory measures that can adapt to the rapidly changing landscape of AI innovation.

In writing this literature review, we aim to provide a timely update on the laws and regulations, along with their salient features, that have been implemented in recent years. We acknowledge that the rapidly evolving nature of AI technologies means that the landscape of AI ethics and regulation is continuously changing. As such, this review is not intended to be an authoritative or

exhaustive account of all existing frameworks and policies, or their effectiveness. Instead, it seeks to offer a concise yet informative overview that highlights the most significant developments in the field. By focusing on recent changes and trends, we hope to create a resource that is both accessible and relevant for readers who are navigating the complex world of AI ethics and regulation.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

(*WEBSITE*) , in 2018, concluded that governance, policy, and regulatory issues related to AI are still in the infancy stage, highlighting two key initiatives on AI regulation, specifically, the "White House Future of Artificial Intelligence Initiative" in 2016, and the Ethics and Governance of Artificial Intelligence Fund have started to "catalyze global research that advances AI for the public interest, with an emphasis on applied research and education in 2017, joint project of the MIT Media Lab and the Harvard Berkman-Klein Center for Internet and Society.

(*[Artificial Intelligence and Machine Learning: Policy Paper]*) covers the challenges that need to be taken into consideration in the development of AI systems. It also developed a set of Principles and recommendations that should be followed in the deployment of AI services on the Internet.

(Heo) covers the implications of AI in settings such as the Internet, the automotive industry, and the healthcare industry, including the benefits and the possible destructive implications. Salient destructive implications discussed include extreme polarization, loss of data privacy, destruction of trust between consumers and providers, unemployment, and controversial moral resolutions.

(*[Allan Dafoe]*) explores how humanity can best navigate the transition to advanced AI systems, focusing on political, economic, military, governance, and ethical dimensions. This research is motivated by the potential for advanced AI to radically transform human welfare, wealth, or power in a way comparable to the nuclear or industrial revolutions

(Marda) highlights the efforts of the Union Ministry of Commerce and Industry in India towards establishing AI governance and regulation. The paper details the formation of the Artificial Intelligence Task Force in August 2017, which identified sectors for widespread application of AI, including manufacturing, financial technology (FinTech), agriculture, healthcare, technology for the differently-abled, national security, environment, public utility services, retail and customer relationships, and education. It further emphasizes the recommendation to establish a nodal agency, the National Artificial Intelligence Mission, to coordinate AI-related activities in India. However, the analysis in the paper points to significant shortcomings in the Task Force's approach, citing a lack of substantial legal, policy, and civil society engagement, which has resulted in an inadequate ethical and social examination of AI's role within the Indian context. Additionally, the Union Ministry of Electronics and Information Technology's initiative, launched in February 2018, to create four committees aimed at exploring AI's impact in areas such as citizen-centric services, data platforms, skilling, R&D, and legal and cybersecurity perspectives, is also mentioned. The paper underscores the need for a more integrated approach that includes ethical, social, and technical considerations to address the limitations and risks associated with data-driven decision-making in AI applications across these sectors.

("Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance") provides a comprehensive analysis of 200 AI ethics guidelines from around the world, examining the principles that are most commonly advocated in these frameworks. The study identifies transparency, justice, non-maleficence, responsibility, and privacy as the most frequently cited principles, echoing findings from previous research but offering broader insights due to its larger and more diverse sample size. It also highlights a significant gap in consensus on how to precisely define these principles, emphasizing the necessity for more practical implementations that can transform these ethical ideals into actionable measures. Furthermore, the study notes an uneven global distribution of AI ethics guidelines, with Europe and North America contributing the majority, a trend possibly influenced by language barriers and database limitations. It reveals that government institutions and private corporations dominate the authorship of these guidelines, likely driven by the tech industry's rapid expansion. Gender disparity also emerges as a critical issue, with men being more frequently represented among the authors, mirroring broader trends in the tech sector. Temporally, the study observes that the peak in the publication of AI ethics guidelines occurred between 2017 and 2019, with transparency increasingly highlighted as a key ethical principle during this period, potentially driven by high-profile cases that brought the need for AI ethics to the forefront. The study also finds that most guidelines are normative and predominantly take the form of recommendations rather than legally binding regulations, indicating a need for more concrete and enforceable actions. It points out a lack of focus on the long-term implications of AI, such as the risk of mass unemployment, suggesting that immediate challenges are perceived as more urgent. Ultimately, the study calls for greater consensus on the definition and implementation of ethical principles and urges attention to AI's long-term societal impacts, while also making its dataset and tools open access to facilitate future research in this area.

[6] provides a comprehensive analysis of 200 AI ethics guidelines from around the world, examining the principles that are most commonly advocated in these frameworks. The study identifies transparency, justice, non-maleficence, responsibility, and privacy as the most frequently cited principles, echoing findings from previous research but offering broader insights due to its larger and more diverse sample size. It also highlights a significant gap in consensus on how to precisely define these principles, emphasizing the necessity for more practical implementations that can transform these ethical ideals into actionable measures. Furthermore, the study notes an uneven global distribution of AI ethics guidelines, with Europe and North America contributing the majority, a trend possibly influenced by language barriers and database limitations. It reveals that government institutions and private corporations dominate the authorship of these guidelines, likely driven by the tech industry's rapid expansion. Gender

disparity also emerges as a critical issue, with men being more frequently represented among the authors, mirroring broader trends in the tech sector. Temporally, the study observes that the peak in the publication of AI ethics guidelines occurred between 2017 and 2019, with transparency increasingly highlighted as a key ethical principle during this period, potentially driven by high-profile cases that brought the need for AI ethics to the forefront. The study also finds that most guidelines are normative and predominantly take the form of recommendations rather than legally binding regulations, indicating a need for more concrete and enforceable actions. It points out a lack of focus on the long-term implications of AI, such as the risk of mass unemployment, suggesting that immediate challenges are perceived as more urgent. Ultimately, the study calls for greater consensus on the definition and implementation of ethical principles and urges attention to AI's long-term societal impacts, while also making its dataset and tools open access to facilitate future research in this area.

The paper (Prem) evaluated over 100 articles on various frameworks, process models, and tools aimed at addressing ethical concerns in the field of AI development. It mainly aims to expand on the work of Morley and colleagues (Morley et al.). It starts off by identifying and collecting a wide range of approaches and conducting a meta-analysis on them.
The meta-analysis conducted in Erich Prem's paper reveals a common structure in the guiding principles of various ethical frameworks. These principles include autonomy (respecting users' decision-making capabilities), beneficence (promoting well-being), non-maleficence (preventing harm), and justice (ensuring fairness), which align closely with traditional bioethical frameworks but are adapted to AI's specific challenges. The primary ethical concerns addressed by these frameworks involve bias in decision-making, a lack of transparency, accountability for decisions, and impacts on user autonomy.
While these frameworks frequently share these ethical pillars, the analysis highlights a significant gap in their depth, particularly regarding real-world implementation. The frameworks often fall short of providing concrete tools or rules for addressing these ethical concerns practically. For instance, the concept of principlism, widely applied in medical ethics, is critiqued as being insufficiently effective for AI ethics due to the absence of specific, actionable guidelines that could guide developers in real-life AI systems. Therefore, despite the broad consensus on the importance of these ethical principles, their application to practical AI development remains underexplored.

Following this, the paper groups all the approaches into the following categories:
Summaries: Overview papers that provide high-level recommendations and case studies to guide ethical AI design.
Notions: Proposals of frameworks, checklists, and metrics to define ethical AI principles, including fairness and explainability.
Procedures: Guidelines, standards, and process models to formalize ethical practices.
Code: Algorithmic methods, design patterns, and software libraries to implement ethical considerations.
Infrastructure: Resources such as datasets and online communities to support ethical machine learning.
Education: Training and educational materials on AI ethics.
Ex-post Assessments: Audits, licenses, and legal frameworks for ethical compliance after development.

The paper then goes on to classify the approaches further based on two methods, classification of approaches by category and classification of approaches by ethical issues addressed, both of which can be seen in Table 7 and Table 8 respectively.

Based on the first table, it is concluded that 26% of the approaches are classified as algorithmic methods,18% focus on conceptual approaches and frameworks, and 12% are software approaches. This means that algorithms and software account for 39% of the approaches. The consensus of the second table reveals that a relatively large group (23%) of the approaches address ethical issues at a general level. Over 50% of the approaches address issues like privacy (19%), fairness and bias (18%), and explainability (16%) while accountability is addressed by only 10% of the approaches. According to the paper, the strong presence of privacy, fairness, and explainability issues in the approaches was already noted by Morley et al. The remaining ethical issues are only addressed in a combined 13% of the approaches.

Hence the classification by category reveals that while technical solutions (accounting for 39%) are a key focus, their effectiveness in addressing broader ethical concerns (e.g., transparency, bias) is limited. The analysis also points out the disproportionate focus on issues like privacy, fairness, and explainability, which receive more attention compared to less frequently addressed but crucial areas like accountability (10%) and transparency.

Finally, the paper highlights several key missing ethical issues in current AI frameworks, such as democratic control and governance, existential threats, environmental costs, political misuse of AI, social cohesion impacts like echo chambers, and the hidden social costs of AI, such as clickworking. These concerns are not easily addressed through technical tools or algorithms. Additionally, the paper emphasizes the lack of practical tools, such as ethics councils, regulatory guides, and accountability labels, and the absence of best practices documentation in AI ethics similar to that found in fields like medicine. AI ethics monitoring, transparency in deployment, and user consent mechanisms also remain underdeveloped. While many tools focus on solving technical ethical challenges, there is a clear need for more comprehensive, systematic approaches that can address these broader societal and ethical concerns effectively.

III. WRITE DOWN YOUR STUDIES AND FINDINGS

Since the publication of Erich Prem's paper in February 2023, several new AI ethics frameworks and guidelines have emerged to address the rapidly evolving landscape of AI technology. These recent initiatives, introduced after the time of the original research, demonstrate significant progress in refining ethical principles and ensuring responsible AI deployment. Among these developments are the finalization of the European Union AI Act (2024), which sets comprehensive regulatory standards for AI across the EU, and the White House Executive Order on Safe AI (October 2023), which outlines a strategic approach to promoting safe AI practices in the United States. Other key frameworks include the NIST AI Risk Management Framework (AI RMF) 1.0, released in January 2023, which provides organizations with risk management strategies for responsible AI, and the OECD Framework on AI Accountability (2023), which focuses on fostering accountability in AI systems. Additionally, global collaborations like the G7 Hiroshima AI Process (May 2023) and national efforts such as the UK's AI Regulation White Paper (March 2023) highlight the growing international cooperation in shaping AI governance.

These new guidelines, including updates from UNESCO, IEEE, and Singapore's government, reflect the urgency to address privacy, transparency, fairness, and accountability in the AI sector. This concise review seeks to build upon Prem's meta-analysis by examining these recent frameworks and also categorizing them based on their approach and ethical focus, and exploring their shortcoming. The paper will first categorize the framework or guideline based on [9][10], give a short synopsis and highlight the key features, and finally discuss whether it addresses the gaps of several other frameworks and guidelines identified in Prem's analysis.

European Union AI Act (2024 Finalization)

- Category: Procedures
- Ethical Issues Addressed: Privacy, fairness, accountability, safety

The European Union AI Act, finalized in 2024, is the world's first comprehensive AI law. The Act categorizes AI systems based on their risk to public safety, security, and fundamental rights. The Regulatory Framework defines 4 levels of risk as

1. Unacceptable risk
2. High Risk
3. Limited Risk
4. Minimal Risk

The AI systems deemed to pose a clear threat, such as social scoring by governments and voice-assisted toys that encourage dangerous behavior, are explicitly banned.

Key elements of the Act include stringent requirements for high-risk AI systems in sectors like healthcare, transportation, and education, ensuring transparency, accuracy, and accountability. The Act also aims to reduce the administrative and financial burdens on small and medium-sized enterprises (SMEs), promoting innovation while safeguarding ethical standards. The legislation sets out clear timelines for implementation: full applicability will be reached by 2026, with certain prohibitions and rules coming into force earlier.

Additionally, the AI Pact encourages voluntary compliance by developers even before the full application of the law, and the newly established European AI Office oversees its enforcement and works closely with member states.

Some key effects of the EU Act include

- Tech companies will be required to label deepfakes and AI-generated content and notify people when they are interacting with a chatbot or other AI system. The AI Act will also require companies to develop AI-generated media in a way that makes it possible to detect.
- The AI Act will set up a new European AI Office to coordinate compliance, implementation, and enforcement. Thanks to the AI Act, citizens in the EU can submit complaints about AI systems when they suspect they have been harmed by one, and can receive explanations on why the AI systems made decisions they did.
- Free open-source AI models that share every detail of how the model was built, including the model's architecture, parameters, and weights, are exempt from many of the obligations of the AI Act.

Gaps addressed:

- Lack of practical tools or implementable steps
  - Clear logical obligations for developers to follow
  - Clear risk-based classification of AI systems with a structured approach to ensure ethical development
- Lack of governmental control and Political Misuse of AI
  - Classifies on the basis of risk and provides stringent rules and obligations to be followed.
  - Citizens can submit complaints about AI systems to the European AI Office
- Lack of labelling
  - Mandates tech companies to label deepfakes and AI-generated content and notify people when they are interacting with chatbots or other AI systems.

Unaddressed issues:

- Environmental issues
- Monitoring and documentation of best practices
- Coaching and consulting

White House Executive Order on Safe AI (October 2023)

- Category: Procedures
- Ethical Issues Addressed: Privacy, fairness, safety, accountability

The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence is a set of guidelines put forth by the White House to regulate and promote responsible AI development in the United States. The primary objective of the executive order is to maintain national security, interests, and international relations while promoting and upholding ethical standards in the development and deployment of responsible AI technologies. The document acknowledges the vast potential of AI to transform industries, enhance national security, and provide economic growth while also emphasizing the need for strict supervision to prevent harmful consequences.

The main areas covered by the document are:

1. National Security and Public Safety: This part of the document focuses on safeguarding AI technologies from misuse, specifically by adversaries. It encourages creating and using AI models in cybersecurity to detect cyberattacks and misinformation. Lastly, it mandates that several rigorous security measures and evaluations must be implemented on AI in critical and sensitive infrastructures.
2. Advancing AI for Economic Growth: This section highlights the potential of AI in economic development by promoting the development of AI systems for sectors such as healthcare, transportation, etc.
3. Civil Rights and Fairness: The major chunk of the executive order is focused on protecting civil rights in the United States. The order ensures fairness in the AI systems used in areas like criminal justice and hiring by mandating rigorous testing and evaluation. It aims to mitigate biases in hiring systems and ensure that minorities are not affected unfairly during the process.
4. AI Safety and Standards: The executive order suggests federal agencies collaborate with the National Institute of Standards and Technology (NIST) to establish standards for transparent, accountable, and safe development and deployment of AI.
5. Workforce Development and Education: This section focuses on preparing the upcoming workforce of the U.S. for the AI boom. It encourages educational programs to include chapters for AI literacy and preparing the future generation to work alongside AI technologies.
6. Ethical Development of AI: Like every other AI ethics guideline or framework, this one too emphasizes the importance of the ethical development of AI systems. It sets forward some rules and guidelines for creating transparent, explainable, and accountable AI systems that align with democratic and ethical values with a great emphasis on the protection of civil rights.
7. International Collaboration: The U.S. recognizes the need for international collaboration in order to develop a flexible yet impactful set of guidelines to follow while developing safe and ethical AI technologies. It therefore encourages collaboration with international partners to develop shared principles and frameworks, especially in terms of security and human rights

Thus the order introduces comprehensive regulations for AI that align with U.S. values, protect civil rights, and national security and promote economic growth.

Gaps addressed:

- Democratic control and governance
    - Encourages international cooperation to develop frameworks to ensure ethical AI development
    - Puts civil rights at the forefront
- Coaching and Consulting
    - Promotes educational institutes and workplaces to include teachings about AI technologies to prepare the future generation to work alongside AI.

Unaddressed Issues:

- Too abstract/lack of actionable points
    - Doesn't provide any obligations or rules to be followed, rather it only sets principles and encourages good practices.
- Environmental impact
- Monitoring and documentation of best practices
- Labeling

NIST AI Risk Management Framework (AI RMF) 1.0

- Category: Procedures
- Ethical Issues Addressed: Fairness, robustness, transparency, accountability

The NIST AI Risk Management Framework (AI RMF) 1.0 is a framework designed to provide private and public sector organizations with a flexible tool to manage risks related to AI systems. The framework is divided on the basis of four core functions: Govern, Map, Measure, and Manage. Like every other framework, it also focuses on the core ethical values of a responsible and accountable AI by promoting transparency, fairness, reliability, and accountability. A significant part of the framework encourages collaboration, where stakeholders across various domains can tailor the framework to their specific needs. This, however, has a negative impact which will be discussed shortly further in this section of the paper.

Unlike several AI ethics frameworks that are often criticized for being highly abstract and lacking actionable steps, the NIST AI RMF, built upon existing principles, takes a more structured approach by breaking down risk management into discrete, implementable steps. It is a practical and detailed guidance provided by NIST on how organizations can integrate ethical values and principles into real-world AI systems. It focuses on mapping and measuring risks, especially in specific applications which offers a sensible and implementable way to manage bias instead of high-level conceptual jargon.

Gaps Addressed:

- Lack of practical tools
    - Provides a structured, actionable set of steps to take (govern, map, measure and manage)

Unaddressed issues:

- Lack of strict consequences
- Too flexible
    - Companies can bend the framework to fit their business agenda rather than focus on ethical development.
    - This is an issue discussed in the paper by Munn [11]

OECD Framework on AI Accountability (2023)

- Category: Procedures
- Ethical Issues Addressed: Accountability, transparency, privacy

The OECD Framework on AI Accountability builds on the OECD AI Principles established in 2019. While the latter emphasized respect of human rights, fairness, transparency and prevention of bias, the framework on AI Accountability turns its focus towards responsible and accountable AI as the name suggests. More specifically, it suggests organizations implement practices that make the AI transparent, explainable, and capable of being

audited. It also outlines guidelines for ethical oversight and risk management to help organizations meet international standards of accountability and compliance. The framework also highlights the need for a rigorous governance structure to manage risks due to the inherent nature and complexity of AI systems.

The OECD framework aligns closely with the themes discussed in Prem's paper, such as the need for accountability, transparency, and the mitigation of bias in AI systems. Like the NIST AI RMF and other frameworks, it focuses heavily on ensuring that organizations are held accountable for their AI systems' outcomes, both in terms of ethical impacts and regulatory compliance. Compared to the high-level principles discussed in Prem's review, the OECD framework adds a layer of accountability, providing mechanisms and strategies for auditing and oversight, thus addressing concerns of transparency and fairness in a more tangible manner. However, while it provides comprehensive guidelines, it may still struggle with the same issue of enforceability that Prem criticizes in other frameworks, especially regarding how these guidelines will be implemented globally across different jurisdictions.

Gaps addressed:

- Accountability and transparency:
    - Promotes responsible AI that is capable of being audited and held accountable
    - Suggests that the company should be responsible for the AI's output
    - Provides strategies for ethical oversight and auditing

Unaddressed Issues:

- Too abstract, no clear actionable steps to implement
- No strict consequences in case of non-compliance

G7 Hiroshima AI Process (May 2023)

- Category: Summaries
- Ethical Issues Addressed: Fairness, societal impact, transparency

Hiroshima AI Process was launched by the G7 under Japan's presidency in May 2023, marking a significant step in fostering international cooperation in the regulation and governance of artificial intelligence. A core goal of the framework is to establish a global consensus on risk management strategies for AI systems, especially in areas where AI poses serious ethical, safety, or societal risks. This entails a focus on topics like AI governance, transparency, auditing, security, and collaboration between the public and private sectors to create shared ethical standards. Furthermore, it nobly advocates for the development of trustworthy AI systems that promote innovation while safeguarding citizens' rights.

Notably, the G7 framework distinguishes itself by promoting international collaboration and addressing the cross-border challenges of regulating AI. However, it also shares some weaknesses with the frameworks discussed by Prem. For example, it remains largely a high-level policy initiative, and like other frameworks, it faces challenges in terms of implementation and enforceability at the national level.

Gaps addressed:

- Democratic control and governance
    - Encourages international collaboration to develop frameworks and guidelines in order to establish global compliance standards.
    - This global scope addresses the concern about the variability of principles, frameworks, and standards across different regions.
- Accountability and transparency
    - Emphasizes the importance of accountability, transparency, and autonomy by placing them at the center of this process
- Strict consequences and governance
    - The process focuses on risk management strategies, especially in areas where AI poses serious ethical safety or societal risks.

Unaddressed Issues:

- Lack of practical tools:
    - This process is a very high-level set of principles offering little to no direction on real-world implementation of the said principles.

UK AI Regulation White Paper (March 2023)

- Category: Notions
- Ethical Issues Addressed: Privacy, fairness, transparency

The UK AI Regulation White Paper, released in March 2023, details the UK's vision for a pro-innovation approach to AI regulation. The paper aims to strike a balance between promoting innovation and ensuring responsible development and deployment of AI.

The White Paper advocates for a flexible regulatory framework where existing regulators take the lead in applying AI-specific principles to their respective sectors. It emphasizes five core principles for AI governance: safety, transparency, fairness, accountability, and contestability. The UK's approach avoids introducing new AI-specific laws and instead supports existing regulatory bodies, while promoting innovation by minimizing the regulatory burden. They are not implementing rigid legislative requirements to avoid hindering AI innovation and to ensure a swift and proportionate response to future innovations. The White Paper also prioritizes public trust and safety, ensuring that ethical considerations are balanced with the benefits of AI.

Just before the AI Summit in October 2023, the UK announced the establishment of the UK Artificial Intelligence Safety Institute (UK AISI). Positioned as the world's first state-backed organization solely focused on advanced AI safety, the UK AISI builds upon the Frontier AI Taskforce's work and is dedicated to conducting fundamental research to ensure public safety in the context of fast-paced AI developments. The UK AISI, while not a regulatory body, will work collaboratively with government and private entities to ensure the UK's AI regulatory approach remains cohesive, evidence-based, and proportionate. Its creation aligns with the principles outlined in the White Paper, ensuring that the UK's AI regulation framework remains adaptable to emerging AI challenges while minimizing regulatory burdens on businesses, especially small and medium-sized enterprises (SMEs).

Gaps addressed:

- Accountability and transparency

Unaddressed Issues:

- Lack of practical tools or implementable steps
- Democratic control and governance
    - Encourages existing regulators to implement AI-specific principles within their respective sectors which may lead to a decentralized and non-rigid set of standards to be followed.
    - This is not inherently a disadvantage, however, it again falls under the trap of being too flexible to ensure rigid and actionable steps for ethical development as discussed in Munn's paper (Munn)

UNESCO Guidelines on AI Ethics Implementation (2023)

- Category: Procedures
- Ethical Issues Addressed: Human rights, fairness, privacy, accountability

The UNESCO Guidelines on AI Ethics Implementation (2023) aim to provide a global, human-centered approach to the governance of AI technologies. These guidelines focus on ensuring that AI development and deployment respect human rights, democratic values, and environmental sustainability. They are built around several key ethical principles: human dignity, fairness, inclusiveness, and the protection of privacy. The guidelines advocate for transparency in AI systems, ensuring that their decision-making processes are explainable and accountable. They also emphasize the need for robust governance frameworks to oversee AI applications in areas like education, healthcare, and labor markets. Importantly, UNESCO stresses the importance of international cooperation, calling for a collaborative global approach to mitigating AI-related risks while promoting benefits.

UNESCO's focus on inclusivity, human rights, and environmental sustainability is in line with the ethical pillars discussed in Prem's analysis. Notably, UNESCO's guidelines emphasize the importance of international cooperation and the need to align AI development with sustainable development goals. In this regard, UNESCO's guidelines aim to bridge some of the gaps identified by Prem, especially in terms of fostering a more integrated, universally accepted framework for AI regulation.

Gaps Addressed:

- Democratic control and governance and strict consequences
    - Emphasizes the importance of international collaboration to develop a unified global approach to AI ethics.
    - Realizes the importance of robust governance to oversee the development of AI in critical areas like healthcare and education.
- Accountability and transparency
    - These guidelines go beyond just the main pillars of ethics to also include a focus on human dignity, inclusiveness, and protection of privacy alongside accountability and transparency.
- Environmental costs
    - Though brief, the guidelines focus on the environmental effects of AI development and deployment.
    - It aims to align some of its principles and practices with the Sustainable Development Goals to ensure the sustainable development of ethical AI systems.

Unaddressed Issues:

- Lack of practical tools and implementable steps

IEEE Ethically Aligned Design (EAD) - Updated Version (2023)

- Category: Procedures
- Ethical Issues Addressed: Fairness, transparency, societal impact

The IEEE Ethically Aligned Design (EAD) is a comprehensive framework aimed at ensuring that AI systems are developed and deployed in a manner that prioritizes human well-being and ethical considerations. The updated version (2023) builds on its predecessors by emphasizing a broad array of ethical principles, including transparency, accountability, privacy, and fairness. The framework encourages stakeholders in the AI ecosystem—ranging from developers to policymakers—to engage in ethical reflection throughout the AI lifecycle. A key aspect of the EAD is its advocacy for interdisciplinary collaboration, promoting the integration of diverse perspectives in the design and governance of AI systems. Furthermore, it calls for the establishment of ethical standards and best practices to guide AI innovation, ensuring that technological advancements do not compromise fundamental human values.

By promoting the inclusion of varied perspectives, the EAD seeks to ensure that ethical considerations are comprehensive and relevant across different contexts. Additionally, the updated version addresses the necessity for accountability mechanisms. However, while the EAD provides a strong theoretical foundation, it shares the common challenge of ensuring that these ethical considerations are effectively translated into actionable practices across the rapidly evolving AI landscape.

The updated EAD effectively addresses several of the shortcomings highlighted in Prem's paper, particularly concerning the need for concrete guidelines and accountability measures. By providing specific recommendations for ethical practices in AI development, the EAD aims to mitigate the ambiguity often associated with ethical considerations in technology. Moreover, its focus on establishing ethical standards and best practices is a direct response to the identified need for more structured frameworks in the AI ethics landscape.

However, like many frameworks, the EAD still grapples with real-world implementation challenges. While it offers a comprehensive

approach to ethical AI design, ensuring adherence to its principles across different organizations and cultures remains a significant hurdle. The EAD's success ultimately depends on the commitment of stakeholders to engage with the framework and implement its recommendations effectively.

Gaps addressed:

- AI governance
- The EAD advocates for interdisciplinary collaboration, promoting a diverse set of perspectives in the design and governance of AI systems.
- The framework aims to ensure that ethical considerations are comprehensive and relevant across different contexts.
- Accountability
- The updated version of the EAD focuses on accountability mechanisms, building on top of an already strong foundation that aims to uphold human autonomy, rights, and values.

Unaddressed Issues:

- Lack of practical tools
    - Even though it provides a comprehensive approach to ethical AI design, it does suffer from the lack of clear directions to implement the principles it proposes for human well-being.

AI Governance Guidelines from the Singapore Government (2023)

- Category: Procedures
- Ethical Issues Addressed: Transparency, accountability, privacy

The AI Governance Guidelines from the Singapore Government, released in 2023, aim to promote the responsible development and deployment of artificial intelligence technologies. These guidelines are designed to assist organizations in implementing ethical AI practices, emphasizing accountability, transparency, and fairness.

Key elements of the guidelines include frameworks for risk management, data governance, and stakeholder engagement. The guidelines also provide practical recommendations for ensuring that the use of AI systems respects privacy and minimizes bias. The Singapore Government aims to establish a robust AI ecosystem that supports innovation and maintains public trust through the development of a collaborative environment involving government, industry, and civil society.

The Singapore Government's AI Governance Guidelines represent a comprehensive effort to merge ethical considerations with practical implementation strategies. The guidelines offer a structured framework that organizations can utilize to navigate the complexities of AI governance.

However, while the guidelines provide valuable direction, their effectiveness hinges on widespread adoption and the ability of organizations to implement these recommendations in diverse contexts. As noted in Prem's critique, the gap between established ethical principles and their implementation remains a challenge for all frameworks. Therefore, continued efforts are needed to ensure compliance and accountability in practice.

Gaps Addressed:

- Governance and democratic control

Unaddressed Issues:

- Lack of actionable steps
- Lack of strict consequences

China's Regulations on AI-Generated Images (2024)

- Category: Notions
- Ethical Issues Addressed: Transparency, accountability, Prevention of Misinformation

China's 2024 Regulations for AI-Generated Images is a draft of guidelines aimed at governing the labeling, transmission, and compliance requirements surrounding AI-generated content. The regulations apply primarily to online service providers that generate or distribute AI-synthesized content, with specific provisions requiring explicit and implicit labeling of AI-generated content, depending on its type (text, image, video, etc.). The goal is to ensure that users are aware of the synthetic nature of AI content, preventing deceptive practices and ensuring transparency. The regulations prohibit the removal or alteration of AI content labels and place significant responsibilities on both service providers and users. Platforms must verify metadata and content to ensure compliance, and violators are subject to penalties according to China's laws.

Compared to other guidelines, such as the EU AI Act, which emphasizes a risk-based categorization of AI systems, China's regulations focus more on public perception, content transparency, and the presentation of AI-generated information. This contrast highlights the different regulatory approaches that governments are taking in the AI ethics space, with China placing a stronger emphasis on information control and the direct regulation of online platforms. This focus might be influenced by concerns over maintaining societal stability and national cohesion, as well as a desire to prevent dissent.

Gaps Addressed:

- Lack of clear actionable guidelines
    - These regulations clearly state what a company or individual needs to do in order to comply with standards.
- Lack of labelling
    - These regulations mandate the need to label AI-generated content appropriately in order to achieve transparency and avoid any sort of confusion.
    - Platforms are expected to verify the meta-data of AI-generated content being uploaded in order to comply with Chinese laws.

- Lack of strict consequences
    - Unlike many regulations and guidelines, these are highly specific and clearly state the repercussions of non-compliance.
    - They provide clear rules on labeling and metadata management, which ensure transparency by requiring AI-generated content to be properly labeled.
    - The regulations strictly prohibit the removal or alteration of AI content labels and place significant responsibilities on both service providers and users.
    - Platforms must verify metadata and content to ensure compliance, and violators are subject to penalties according to China's laws.
- Governance and political misuse
    - By enforcing laws to mitigate the spread of misinformation or deceptive content, China directly addresses the issues of misuse of AI systems.

Unaddressed Issues:

- Broader societal issues
    - Though it is not intended to, the regulations do not focus at all on other pillars of ethical AI like privacy, accountability, fairness and its impact on society.

Examples of Companies Incorporating Mentioned Frameworks or Guidelines

Several prominent companies demonstrate their commitment to ethical AI development and compliance with various international laws, guidelines, and frameworks. Examples of two such companies are given below:

- Microsoft: The RAI Transparency report released by Microsoft in May 2024, explains how they develop responsible generative applications, make decisions about releasing generative applications, and support their customers in building responsibly using their services like Azure OpenAI. The report provides an extensive view into how their Generative AI requirements set by their Responsible AI Standard align with the core functions of the National Institute for Standards and Technology (NIST) AI Risk Management Framework. (*[Responsible AI Transparency Report]*)

- IBM: IBM has long been an advocate of ethical AI and ensures its AI solutions adhere to global standards like ISO/IEC 42001 and NIST AI RMF. IBM has published an AI ethics charter and offers frameworks for AI governance that emphasize fairness, transparency, and accountability. Their AI systems, particularly Watson, are continuously audited to meet both voluntary frameworks and legal obligations, such as the upcoming EU AI Act. IBM also ensures their AI development aligns with the Sustainable Development Goals provided by the UN. ("IBM Welcomes OECD Principles for the Development and Use of AI")

Companies like Google have implemented their own ethical values that share many similarities with the frameworks and guidelines mentioned here. Hence most major companies have realized the importance of ethical development and deployment of AI to ensure the protection of human rights, values, and dignity.

IV. CONCLUSION

In conclusion, the increase in the number of AI ethics frameworks and guidelines after February 2023 indicates the massive increase in awareness of the need for regulation and robust governance structures for AI. Initiatives like the European Union AI Act and the White House Executive Order on Safe AI are influential landmark policies, while standards like NIST's AI RMF 1.0 offer practical risk management tools. These new guidelines emphasize the need for transparency, fairness, and the mitigation of bias in AI systems, values in line with existing ethical principles. The guidelines are however developed to meet challenges presented by rapidly developing AI technologies. The global collaboration represented by bodies such as the OECD, G7, and UNESCO further demonstrates that addressing AI's ethical challenges requires cooperation across borders and sectors, ensuring a more harmonized and responsible future for AI development. By covering such frameworks and regulations, this paper puts a spotlight on the progress made ineffective regulations and frameworks, and also the ongoing necessity for adaptive, interdisciplinary approaches to AI ethics in the face of emerging ethical dilemmas.

REFERENCES

[1] Heo, Jung Hwan. "Ethical Review in The Age of Artificial Intelligence." *The AI Ethics Journal*, vol. 2, no. 2, 2021, https://doi.org/10.47289/AIEJ20210716-4.

[2] "IBM Welcomes OECD Principles for the Development and Use of AI." *IBM Policy*, 22 May 2019, https://admin03.prod.blogs.cis.ibm.net/policy/oecd-principles-ai/.

[3] Marda, Vidushi. "Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Nov. 2018, https://doi.org/10.1098/rsta.2018.0087.

[4] Morley, Jessica, et al. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics*, vol. 26, no. 4, Dec. 2019, pp. 2141–68.

[5] Munn, Luke. "The Uselessness of AI Ethics." *AI and Ethics*, vol. 3, no. 3, Aug. 2022, pp. 869–77.

[6] *[Artificial Intelligence and Machine Learning: Policy Paper]*. https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf. Accessed 15 Oct. 2024.

[7] Allan Dafoe. https://cdn.governance.ai/GovAI-Research-Agenda.pdf. Accessed 15 Oct. 2024.

[8] Responsible AI Transparency Report. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1l5BO. Accessed 15 Oct. 2024.

[9] Prem, Erich. "From Ethical AI Frameworks to Tools: A Review of Approaches." *AI and Ethics*, vol. 3, no. 3, Feb. 2023, pp. 699–716.

[10] *Website*. https://www.researchgate.net/publication/325934555_Artificial_Intelligence_A_Study_on_Governance_Policies_and_Regulations.

[11] "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance." *Patterns of Prejudice*, vol. 4, no. 10, Oct. 2023, p. 100857.

AUTHORS

First Author – Roshan T Baby, BTech CSE, D.Y. Patil International University, Akurdi and 20210802003@dypiu.ac.in.
Second Author – Virenn Vaatts Jay, BTech CSE, D.Y. Patil International University, Akurdi and 20210802170@dypiu.ac.in