Standards, frameworks, and legislation for artificial intelligence (AI) transparency

Article *in* AI and Ethics · January 2025

DOI: 10.1007/s43681-025-00661-4

7 authors, including:

Brady D. Lund
University of North Texas
223 PUBLICATIONS **4,142** CITATIONS

SEE PROFILE

Ravi Varma Kumar Bevara
University of North Texas
13 PUBLICATIONS **82** CITATIONS

SEE PROFILE

Standards, Frameworks, and Legislation for Artificial Intelligence (AI) Transparency

Brady D. Lund, Zeynep Orhan, Nishith Reddy Mannuru, Ravi Varma Kumar Bevara, Brett Porter, Meka Kasi Vinaih, and Padmapadanand Bhaskara

Abstract

The global landscape of transparency standards, frameworks, and legislation for artificial intelligence (AI) shows an increasing focus on building trust, accountability, and ethical deployment. This paper presents comparative analysis of key frameworks for AI transparency, such as the IEEE P7001 standard and the CLeAR Documentation Framework, highlighting how regions like the United States, European Union, China, and Japan are addressing the need for transparent and trustworthy AI systems. Common themes across these standards include the need for tiered transparency levels based on system risk and impact, continuous documentation updates throughout the development and revision processes, and the production of explanations tailored to various stakeholder groups. Several key challenges arise in the development of AI transparency standards, frameworks, and legislation, including balancing transparency with privacy, ensuring intellectual property rights, and addressing security concerns. Promoting adaptable, sector-specific transparency regulatory structures is critical in the development of frameworks flexible enough to keep pace with AI's rapid technological advancement. These insights contribute to a growing body of literature on how best to develop transparency regulatory structures that not only build trust in AI but also support innovation across industries.

Introduction

The Biden-Harris Administration's Executive Order 14110, known as the order on “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” directs U.S. government agencies to ensure the development of AI that is safe, reliable, and transparent. In response to this order, U.S. government agencies have spearheaded efforts to produce new policies and educational initiatives that address the potential risks posed by AI technologies, while simultaneously utilizing their positive applications. Concurrently, many politicians have increasingly focused on regulating and overseeing the use of emerging AI technologies to safeguard against misuse and unintended consequences (Bareis & Katzenbach, 2022). Jobin et al. (2019) note that this growth in interest in safe, reliable, and transparent AI through standards, policies, and legislation is mirrored in many other countries and regions throughout the world.

The rapid rise of generative AI has triggered accelerated discussions about how these technologies should be managed responsibly by developers. Central to this debate is the issue of transparency in AI systems, including these important questions: Should AI-generated content be watermarked? Should users be informed when an AI system is in use? Should the inner workings of these systems be disclosed? Should the data used to train AI models be made public? These questions highlight a fundamental challenge in AI development, providing proper information sharing about how these technologies work. This level of transparency is critical as AI continues to influence key sectors such as healthcare and information systems, as transparency is essential for maintaining public trust, fairness, and ensuring the responsible use of these technologies (Felzmann et al., 2020; Reinhardt, 2023).

Transparency for AI refers to providing clear information about how these systems operate, what data they rely on for training, and the reasoning behind their decision-making processes (Larsson & Heintz, 2020). Transparency is important both for upholding ethical and legal standards as well as for building trust with users, as people are more likely to have confidence in AI systems when they understand how these systems generate results and make decisions (Von Eschenbach, 2021). Furthermore, transparency is necessary to address and correct mistakes in AI systems. As AI often handles sensitive personal data when making decisions, ensuring privacy, proper authorization, and adherence to ethical standards is vital. These decisions can have a profound impact on individuals' lives. It is imperative that users are well-informed about the data AI systems consume, how that data is processed, and the outcomes these systems produce.

This paper examines the status of standards (examples or models established by an authority), frameworks (basic conceptual structures proposed by an individual or group), and legislation (legal rules established and enforced by a governing body) for AI transparency around the world, taking a narrower and updated view of this specific issue that pairs with broader analyses of global AI policy such as Hickok (2021). Presently, little is known about the status of AI transparency regulatory structures around the world, where this awareness could be critical to ensuring consistency and quality across all jurisdictions and potential applications of emerging

AI technologies. Given the unique position of the United States within the landscape of AI development – being home to major AI developers Nvidia, Google, Microsoft, OpenAI, and Tesla, among others – the first major section of this paper is dedicated to the status of AI transparency standards within the U.S. This is followed by a discussion of major efforts related to AI transparency guidelines from around the world. We will then conclude by discussing the overlap in discrepancies among standards and their implications for AI developers, users, and researchers.

The Status of AI Transparency Standards, Frameworks, and Proposed Legislation in the United States

Owing to its founding during the Age of Enlightenment, the American people generally adhere to a set of moral and political philosophies emerging from this period, such a Social Contract theory, which provide a foundation for much of the leading ethical theory applied to AI development today (Srinivasan & Ghosh, 2023). The Social Contract promotes a dynamic between individuals and institutions (such as AI developers) whereby the individual puts their trust in institutions and voluntary relinquish some rights in exchange for the services provided by the institution, which in turn is expected to be accountable to the individual. Transparency is a means through which this Social Contract can be enforced in terms of AI development. The user of the AI tool acknowledges that they are forfeiting some rights (sharing data, storing data on the developer's servers) and the developer acknowledges that they will be responsible to the user (by keeping data private and secure). By being transparent about how the user's data will be stored and used, the developer can maintain the trust of the user. These principles underpin the development of many of the major transparency standards, frameworks, and legislation that have been proposed today.

In this section, we focus on the status of several AI transparency standards, frameworks, and legislation within the United States, a country that has long been on the leading edge of technology development including in artificial intelligence. Five proposals are reviewed, noting the basic tenants of the proposal and the potential positive and negative outcomes of their implementation on AI development: the National Artificial Intelligence Advisory Committee (NAIAC) AI Transparency Standards, the IEEE Standards for Transparency of Autonomous Systems, the U.S. Department of Health and Human Services HTI-1 Final Rule, the CLeAR Documentation Framework for AI Transparency, and the proposed U.S. Research, Innovation and Accountability Act.

NAIAC AI Transparency Report

Transparency in the development and deployment of AI has emerged as a cornerstone of ethical AI governance (Memarian & Doleck, 2023). A recent analysis of the codes of conduct from 16 companies across diverse sectors revealed that 14 of these organizations explicitly prioritize transparency in their AI ethics policies (Balasubramaniam et al., 2023). This finding emphasizes transparency's important role in providing trust and accountability within AI systems (Binns, 2018).

The link between transparency and trust is well-documented, particularly in guaranteeing that AI operations are comprehensible to users. According to the surveyed organizations, the primary motivations for emphasizing transparency include building and maintaining user trust, improving security, and facilitating the evaluation of system efficiency (NAIAC, 2024). These insights align with extensive research indicating that transparency is indispensable for establishing trust in AI, particularly in high-stakes domains such as healthcare and finance (Rudin, 2019).

Towards Standards for Data Transparency for AI Models

NAIAC is a group of prominent figures in industry, academia, and labor organizations who were brought together to investigate the potential for creating baseline standards for data transparency among the developers of AI models. Their goal was to propose what could be considered minimum acceptable standards for transparency, acknowledging that the diverse stakeholder groups involved in AI development and deployment would likely have considerable disagreement about the extent to oversight is needed in AI.

The NAIAC (2024) looked to existing proposed and implemented standards in specific disciplines, such as healthcare, to guide them. The committee notes in their publication the need to balance competing tensions between transparency, privacy, and intellectual property rights. Stakeholders disagree as to the extent to which one of these interests should be prioritized over the others. Industry leaders in AI development may wish to protect their intellectual property (IP) over all other considerations, whereas academics and regulators may seek to balance transparency and privacy, even if it requires sacrificing some IP rights. The greatest outcome of the NAIAC is perhaps the fact that considerable give-and-take will be necessary to establish AI transparency standards that will be amendable to all stakeholder populations.

IEEE Standards for Transparency of Autonomous Systems (2021)

The accelerated growth of innovation with AI and autonomous systems, including systems like self-driving cars, drones, and AI-powered medical devices, has produced major advancements across industries in a short period of time. However, these developments have also prompted new concerns about the transparency and reliability of these systems (Winfield et al., 2021). Coeckelbergh (2020) identifies AI transparency as a critical factor in combatting the potential risks associated with autonomous systems operating as “black boxes.” These autonomous systems often lack clear information about their decision-making processes, making it difficult for users to assess the systems’ safety and reliability. This opacity about systems’ operations has fueled skepticism and fear. This trepidation is particularly pronounced in high-stakes domains like autonomous vehicles and healthcare, where decisions can have life-altering consequences.

In response to these emerging challenges, the Institute of Electrical and Electronics Engineers (IEEE) Standards Association introduced the IEEE P7001 standard in 2021, providing a structured framework to ensure transparency in AI-driven autonomous systems (IEEE Standards

Association, 2022). The standard aims to address the lack of clarity surrounding how these systems operate and make decisions by establishing measurable and verifiable transparency benchmarks. Through these guidelines, the standard seeks to enhance trust and accountability while enabling effective regulation and oversight, ensuring that transparency requirements are tailored to varying levels of system complexity and stakeholder needs (Winfield et al., 2021; Theodorou et al., 2017).

The Role of Transparency in Trust and Accountability

Transparency plays a vital role not only in providing user trust but also in ensuring accountability, particularly in environments where autonomous systems operate in sensitive or life-critical capacities. Pagallo (2017) highlights the importance of legal frameworks, such as the General Data Protection Regulation (GDPR) from the European Union, in ensuring accountability for autonomous systems, emphasizing the role of secondary rules in addressing risks and supporting transparency. High-profile incidents involving autonomous vehicles, where opaque decision-making processes have led to accidents, have further eroded public trust and highlighted the need for standardized transparency (Goodall, 2014).

The IEEE P7001 standard addresses these concerns by adopting a multi-tiered approach designed to meet the diverse transparency requirements of various stakeholders, including end-users, safety certifiers, and investigators. This multi-stakeholder framework ensures that transparency is implemented comprehensively, offering basic user-friendly explanations for non-expert users while providing detailed technical audits for expert stakeholders (Winfield et al., 2021).

A Multi-Tiered Framework for Explainability

The IEEE P7001 standard outlines a tiered framework for explainability, which identifies the unique transparency needs of distinct stakeholder groups. Non-expert users, such as passengers in autonomous vehicles or patients using AI-driven medical systems, require simple, accessible explanations that clarify system behavior without overwhelming them with technical details. These explanations help users understand the rationale behind decisions, reducing the perception of AI systems as incomprehensible "black boxes" and fostering confidence in their use (Winfield et al., 2021).

Conversely, expert stakeholders—including developers, certifiers, and investigators—require in-depth technical insights to ensure the safety, reliability, and accountability of these systems. This group needs access to detailed system logs, architectural documentation, and decision-making processes to certify compliance and investigate incidents. By providing such comprehensive transparency, the standard supports rigorous auditing and accountability, ensuring that autonomous systems meet stringent safety and performance criteria (IEEE Standards Association, 2022).

Transparency Levels and Compliance Mechanisms

The IEEE P7001 standard introduces a comprehensive framework with defined transparency levels, ranging from basic to advanced, to accommodate the diverse needs of users and regulatory bodies (Winfield et al., 2021).

Level 1: Basic transparency, including user manuals and simple system explanations for consumer products.

Level 2: Interactive materials or system descriptions tailored to non-expert users.

Level 3: Real-time explanations and access to system logs for expert stakeholders.

Levels 4 and 5: Detailed and continuous transparency, granting full access to system behavior, decision-making processes, and training data for high-risk applications.

This tiered structure provides flexibility, allowing systems to scale their transparency measures based on their complexity and associated risks. Lower levels of transparency suffice for low-risk applications, while higher levels are mandatory for critical systems like autonomous vehicles or medical diagnostic tools. Research by Tsamados et al. (2021) highlights the importance of tailored approaches to transparency, noting that varying transparency requirements for different stakeholders can address ethical challenges and support accountability.

Flexibility and Scalability

P7001's compliance mechanisms are designed to evolve alongside technological advancements, guaranteeing the standard remains relevant as AI systems become more sophisticated. Morley et al. (2020) highlight the need for ethical frameworks to evolve alongside the increasing complexity of AI technologies, ensuring practical tools and methods remain effective in addressing emerging challenges. P7001 encourages innovation while ensuring that high-risk systems are held to rigorous transparency standards without overburdening simpler applications by providing scalable transparency solutions.

The IEEE P7001 standard represents a critical advancement in the governance of AI technologies. Its structured framework facilitates the objective evaluation of transparency, establishing clear benchmarks that benefit both developers and regulatory bodies. This system emphasizes continuous improvement, maintaining transparency as a primary focus in consumer-facing and high-risk applications (IEEE Standards Association, 2022).

Developers benefit from P7001 through its provision of clear guidelines for achieving suitable levels of transparency, striking a balance between innovation and ethical accountability. Regulators are equipped with a robust tool to assess system transparency, taking into account the associated risk and complexity, thereby supporting accountability and legal compliance

(Winfield et al., 2021). The IEEE P7001 standard serves as a bridge between technological progress and public trust. Clear and comprehensible explanations of AI decision-making processes empower non-expert users, building confidence in these systems. Simultaneously, expert stakeholders gain tools that facilitate thorough evaluation and accountability. This approach establishes AI technologies as powerful and innovative while emphasizing their transparency, responsibility, and trustworthiness (Winfield et al., 2021).

U.S. Department of Health and Human Services HTI-1 Final Rule

AI has taken a major role in healthcare, including within the areas of clinical decision-making, diagnosis, and patient management (Yekaterina, 2024). AI-driven tools offer the potential to enhance care quality, improve efficiency, and reduce human error. However, their deployment raises major concerns due to the sensitive nature of personal health data involved. This data is safeguarded by strict privacy regulations, which must be taken into consideration when developing systems that integrate this data. Managing protected health information brings ethical, privacy, and transparency issues to the forefront. In this context, transparency is crucial for building trust among patients and healthcare providers, since these tools often influence critical health decisions.

The Challenges of Transparency in Healthcare AI

Healthcare AI faces notable transparency challenges, particularly due to the "black box" nature of many AI models. These systems often deliver predictions or decisions without clear explanations of how they were derived. In healthcare, such opacity can undermine trust, leaving both patients and providers uncertain about the AI's reliability and accuracy, which can lead to harmful outcomes for the patient (Fernandez-Quilez, 2023).

Another pressing challenge is bias in AI models, which stems from the data on which they are trained. Since AI systems rely heavily on the quality of their training data, biased or incomplete data can perpetuate existing disparities in healthcare outcomes. Transparency in data sources and algorithms is essential to identify and eliminate these biases. Healthcare providers must be able to clearly explain AI-driven decisions, emphasizing the importance of transparent processes in handling and interpreting patient data.

Advancing Transparency: The HTI-1 Final Rule

The HTI-1 Final Rule represents a significant step in the U.S. Department of Health and Human Services' (HHS) broader effort to advance healthcare technology (Everson et al., 2024). Building on the 21st Century Cures Act, this rule seeks to enhance transparency, interoperability, and patient access to health information. Specifically, it establishes key criteria for using health IT, including AI-driven tools, to ensure health data is accessible and actionable across different systems. The rule aims to reduce the opacity of AI models while maintaining patient privacy by

promoting the use of APIs and requiring the disclosure of data sources and decision-making processes.

Interoperability Requirements

The HTI-1 Final Rule strengthens interoperability requirements to ensure seamless data exchange between healthcare systems. The rule allows AI models to access comprehensive, up-to-date data from diverse platforms by mandating that health information be shared in standardized, accessible formats. This interoperability enhances the accuracy of AI predictions, enabling systems to draw insights from multiple data sources.

Patient Access to Health Information

A cornerstone of the HTI-1 Final Rule is its emphasis on improving patients' access to their own health information. Under this rule, patients must have secure, barrier-free access to their electronic health information. The rule empowers patients to choose third-party applications and tools to retrieve and review their health data in real time through APIs. This provision not only provides transparency but also promotes patient engagement in their healthcare journey.

Use of Application Programming Interfaces (APIs)

The rule requires healthcare organizations, developers, and certified health IT vendors to implement standardized APIs, facilitating third-party access to health data. These APIs enhance interoperability and ensure seamless data sharing. Notably, the rule mandates that APIs enable patients to access their data directly, without relying on third-party applications, while also preventing information blocking.

CLeAR Documentation Framework for AI Transparency

The CLeAR (Comparable, Legible, Actionable, and Robust) Documentation Framework, introduced by scholars associated with Harvard Kennedy School's Shorenstein Center and Microsoft Research, provides a systematic method to improve transparency in AI systems (Chmielinski et al., 2024). This framework responds to the escalating demand for accountability and comprehension in the development and deployment of AI, especially as these technologies gain prominence across diverse sectors.

The principles of the framework support AI data that is consistent across systems, understandable to the target audience, usable, and sustainable in the long term. The importance of these concepts lies in their ability to provide trust in the development of AI, to make informed decisions, and to encourage responsible behavior (Gebru et al., 2021; Mitchell et al., 2019). Data will be useful in analyzing and comparing data on various AI applications. This principle supports a level of design that supports meaningful comparisons while recognizing differences in AI applications. For example, the documentation for an AI system utilized in healthcare diagnostics will differ from that of an email spam detection system; however, both should

conform to specific common elements to facilitate relevant comparisons (Chmielinski et al., 2024).

Legibility emphasizes the importance of ensuring that AI documentation is easily understandable for the target audience. This principle recognizes that diverse stakeholders, including technical experts, policymakers, and the general public, may necessitate different levels of detail and explanation (Hind et al., 2020). The framework posits that documentation must explicitly define the system's objectives, functionalities, and constraints, customizing the content to align with the distinct requirements and knowledge levels of various user demographics. Dodge et al. (2019) emphasize the role of explainable AI in fostering fairness judgments, highlighting how different explanation styles can impact users' perceptions and understanding of machine learning systems.

The actionable component of the framework highlights the practical applicability of the documentation. It is essential to furnish stakeholders with adequate information to facilitate informed decision-making regarding the utilization, assessment, or governance of the AI system (Stoyanovich & Howe, 2019). This may encompass information regarding the performance metrics of the system, possible biases, or recommendations for suitable deployment contexts. The concept of robustness in documentation pertains to its capacity to maintain relevance and accuracy throughout its lifespan. Considering the evolving landscape of AI development, the framework emphasizes the necessity for frequent updates and the implementation of procedures to ensure documentation is maintained throughout the lifecycle of an AI system (Hutchinson et al., 2021).

Implementing the CLeAR framework necessitates careful consideration of multiple trade-offs. For instance, the pursuit of comparability may occasionally clash with the necessity for tailored solutions in specific AI applications. In a similar vein, achieving legibility for a diverse audience may necessitate the simplification of technical details, which could compromise the depth of information sought by specialists (Chmielinski et al., 2024). The framework is constructed upon and is consistent with earlier documentation initiatives within the discipline, including datasheets for datasets, model cards, and various tools aimed at enhancing transparency (Holland et al., 2018; Pushkarna et al., 2022). The objective is to establish a cohesive methodology that can be customized for various AI domains and applications.

The CLeAR framework's acknowledgment of the sociotechnical aspects inherent in AI systems stands out as one of its primary strengths. It promotes the importance of documenting not only technical specifications but also the wider context surrounding the development and deployment of AI systems. This includes details regarding the data utilized for training, the methodologies employed in system design, and the prospective societal implications (Bender et al., 2021). The framework emphasizes the need to fully integrate knowledge throughout the AI lifecycle, not as an afterthought helping to identify and mitigate potential issues early in development, which can reduce the cost and complexity associated with post-deployment solutions (Richards et al., 2020).

The CLeAR framework requires a firm commitment and appropriate allocation of resources. The authors acknowledge that creating effective and efficient knowledge is a labor-intensive process that will require the use of new tasks and working methods. Nonetheless, they contend that the advantages related to enhanced transparency, accountability, and trust surpass these associated costs (Madaio et al., 2020).

A Look at the U.S. Research, Innovation and Accountability Act

The U.S. Research, Innovation and Accountability Act is a piece of legislation proposed by a bipartisan group of prominent legislators, including: John Thune (R-SD), Amy Klobuchar (D-MN), Roger Wicker (R-MS), John Hickenlooper (D-CO), Ben Ray Luján (D-NM), Shelley Moore Capito (R-WV), and later cosponsored by Tammy Baldwin (D-WI) and Cynthia Lummis (R-WY). As of September 2024, the bill passed the Senate Committee on Commerce, Science, and Transportation and was awaiting introduction to the full Senate. This piece of legislation is significant and a departure from the standards discussed above in that it would legally compel AI developers to follow some standards, rather than rely on a pledge or simply provide a framework. As such, the passage of this legislation would represent a considerable step toward AI transparency reform.

This legislation is a comprehensive mandate to reign in the unfettered and, at times, unethical development of AI models, by placing certain constraints and oversight over developers and deployers of these models. Specifically, the legislation incorporates several stipulations relevant to these entities:

- Internet platforms must disclose when AI is in use. Presently, it is feasible to navigate through a website and interact with several different specialized AI models without any awareness that you are doing so. You may be using AI when you conduct an information retrieval task, message customer service, or choose a new article to read based on your prior viewing behavior. Users will need to be informed when AI is integrated into these tools and have access to explanations of how it is used.
- The U.S. federal government will regulate “high-impact AI systems,” defined as systems, other than those designed for national defense purposes, that make decisions that have a significant effect on access to housing, employment, credit, education, health care, or insurance. These systems have considerable effect on the livelihood and wellbeing of the public. In an unregulated environment, AI may be allowed to make discriminatory decisions when approving or declining housing applications, determining who to hire for an open job position, making decisions on admissions to prestigious universities, and offering insurance to higher risk individuals and households (Attard-Frost et al., 2023). In the current environment, these decisions could be made by AI without the deployers even needing to inform those impacted by this usage.
- Under this legislation, deployers of high-impact AI systems would be required to submit annual design and safety plans to National Institutes of Standards and Technology

> (NIST). These reports would be required to include several key pieces of information. The developers would need to specify the purpose of the AI system, its intended use cases, and the deployment context. In other words, they would have to explain why the system is needed, how it would be used, and for what specific purposes, to ensure there is no abuse. Furthermore, the developers would need to explain the benefits of the system. They would also need to note what data was used to train and operate the model, how it is evaluated (and an evaluation schedule), and what steps will be taken to monitor the system and ensure it is not misused (Artificial Intelligence Research, Innovation, and Accountability Act of 2023, 2024).

Under this legislation, assessment criteria would be used to evaluate the systems. These criteria would ensure that the value of the system outweighs the risks posed. This oversight would be carried out by the Under Secretary of Commerce for Standards and Technology in accordance with the NIST. Revised standards would be established by an AI certification advisory committee. If a developer or deployer is found to have an AI system that is not in compliance with the standards, they must take immediate remedial action to address the issue or face penalties of up to $300,000 USD and termination of the system (U.S. Research, Innovation and Accountability Act of 2023, 2024). Additional civil penalties could also be imposed if awareness and intent to use non-compliant systems is uncovered.

This legislation also provides additional support in educating the public about issues related to AI and transparency education through the establishment of an Artificial Intelligence Consumer Education working group comprised of individuals from institutions of higher education, AI developers, and representatives of various sectors and industries. The initiatives of this working group could lead to additional legislation aimed at advancing AI education and addressing the growing public demand for transparency in AI systems, which several authors emphasize as essential for safeguarding users of these technologies (Memarian & Doleck, 2023; Schiff, 2022). Miller (2019) emphasizes that explainability in AI is a key component for fostering user trust and understanding, suggesting that effective explanations tailored to user needs are critical for responsible AI deployment.

The efforts outlined in the AI Research, Innovation, and Accountability Act are a substantial step forward in U.S. AI law and policy. Previously, there have been efforts to pass comprehensive AI transparency regulations for the European Union (Wulf & Seizov, 2020) and China (Shin, 2019), but the United States has been left woefully behind. While this lack of legislation may entice AI developers to center their development and AI offering in the country, it also leaves the people of that country vulnerable to manipulation and lapses in privacy and security. It is possible that, as this legislation advances, other standards mentioned in prior sections could play some role or even be directly adopted into the law. The current legislation mentions the NIST as a policy and standards leader. The NIST already has several existing standards for AI explainability and risk

management (Swaminathan & Danks, 2024; Quinn et al., 2020). The organization is a logical leader for transparency standards and enforcement as well.

The Status of AI Transparency Standards, Frameworks, and Legislation Around the World

This section of the paper reviews the status of AI transparency standards, frameworks, and legislation among several regions around the world and multinational organizations: Africa, Canada, China, European Union, India, Japan, OECD, Russia, and ASEAN. We note the status of standards and their basic tenets, as well as the potential impact on developers and users within each of these regions.

AI Regulations in African Countries

African nations are gradually formalizing AI governance frameworks to promote ethical and responsible deployment of AI, aligning with international standards while addressing their unique needs and priorities. Okolo (2024) emphasizes the urgent need for AI governance across Africa, where only seven countries—Benin, Egypt, Ghana, Mauritius, Rwanda, Senegal, and Tunisia—have drafted national AI strategies, with none yet implementing formal AI regulations. While AI-specific policies are limited, 36 out of 54 African nations have established data protection laws, providing a basis on which AI regulation could be built. Drawing from the European Union's experience, where the GDPR laid the foundation for the upcoming EU AI Act, Okolo argues that data governance in Africa could be expanded to cover data quality, transparency, privacy, and labor protections for data workers. These adjustments would need to account for Africa's unique socio-economic challenges, creating frameworks that are culturally and regionally relevant.

Additionally, Okolo advocates for a collaborative approach to AI governance, emphasizing cross-sector engagement to provide responsible AI development and effective oversight. Okolo points to recent initiatives in the United States and United Kingdom, such as the U.S. Executive Order on AI and the UK's AI Safety Summit, where broad stakeholder involvement from advocacy groups, academia, policymakers, and tech companies has driven AI governance. For Africa, Okolo suggests that local stakeholders—civil society, industry leaders, and academic experts—should play active roles on advisory boards and expert panels, following models seen in United Nations and OECD-led efforts. This inclusion would promote better data stewardship and ensure AI governance frameworks align with Africa's specific cultural and socio-economic realities, supporting equitable AI implementation and contributing to global advancements in ethical AI governance.

Australia: Artificial Intelligence Ethics Framework

Australia's approach for managing the ethical challenges of AI is encapsulated in the Artificial Intelligence Ethics Framework, which was introduced to guide organizations in the responsible development and deployment of AI technologies. This framework is not legally binding but

serves as a set of voluntary guidelines that encourage companies to adopt key ethical principles such as fairness, accountability, and transparency. These principles are designed to help organizations ensure that AI systems are used in ways that are beneficial and just, promoting trust among the public and other stakeholders. The framework envisions an environment where AI innovations can thrive while guaranteeing that they contribute positively to society and do not exacerbate inequalities or harm vulnerable populations (Australian Government, 2021).

Although the framework is voluntary, it has had a significant influence on how businesses and developers approach AI in Australia. The framework helps organizations navigate the complex ethical landscapes that AI technologies often present by providing a clear set of guidelines. It encourages a proactive approach to ethical considerations, from the design phase through to deployment, targeting that AI systems are not only technologically effective but also socially responsible. The adoption of these guidelines is seen as a step towards enhancing Australia's reputation in ethical AI development globally, aligning with international standards and contributing to global discussions on AI governance (Australian Government, 2021).

Canada: Directive on Automated Decision-Making

Canada's Directive on Automated Decision-Making, implemented in 2021, represents a significant governmental initiative to regulate AI systems within public administration. This directive mandates federal agencies to assess the impact of AI systems to ensure they are used responsibly in public decision-making processes. The primary objective is to assess the potential impact of these systems on the public and to prevent them from perpetuating bias or causing harm. The directive sets out clear guidelines for transparency, accountability, and fairness in the deployment of AI technologies, focusing on protecting citizens' rights and promoting trust in government use of AI (Government of Canada, 2021).

Federal departments are required to conduct algorithmic impact assessments for all AI deployments, regardless of their scale. These assessments are crucial for identifying potential risks and biases associated with automated decision systems. The directive also mandates detailed documentation of the design and deployment processes of AI systems, enhancing the transparency of how decisions are made and providing a basis for accountability. It serves as a model for both public and private sector AI governance by systematically addressing the ethical and legal challenges associated with AI. It emphasizes Canada’s commitment to guaranteeing that AI technologies are implemented in a manner that respects privacy, human rights, and the rule of law (Government of Canada, 2021).

China: Artificial Intelligence Governance Principles

In an interview with Charis Liu, Ngor Luong (2024) discusses China’s strategic approach to AI governance, particularly in its interactions with the global South. China’s framework comprises four main elements: a “people-centered” approach (*yi ren wei ben*), a focus on national sovereignty to counter disinformation and maintain digital sovereignty, standards for risk

assessment in AI systems, and increased representation for developing countries in global AI governance through organizations like the United Nations. Despite these stated goals, skepticism remains due to the contrast between China's domestic policies, which involve stringent information control, and its outward call for collaboration.

China's influence in the global South is further established through platforms like the Digital Silk Road, where it aids in AI infrastructure and talent development, especially in Southeast Asia and Latin America. The Shanghai Declaration on Global AI Governance supports this by promoting technical training, secure data-sharing practices, and cultivation of AI talent, positioning China as a leader in AI education and innovation in emerging economies. While China claims to manage risks like cybersecurity and misuse of AI, Luong highlights concerns, such as Huawei's 5G vulnerabilities, that suggest a prioritization of surveillance over transparency.

Luong emphasizes that China's multilateral platforms, including the Digital Silk Road and the Belt and Road Initiative, serve to further its digital sovereignty narrative, particularly in areas less influenced by the U.S. China appeals to governments focused on economic progress by advocating for AI's role in fields like industrial innovation and smart city development. However, Luong advises the global South to carefully evaluate China's "Ethical Norms for New Generation AI," which emphasizes values aligned with China's political agenda rather than universally recognized ethical principles. This cautious approach allows countries in the global South to weigh the implications of China's AI governance on their own development and ethical standards.

Europe: General Data Protection Regulation (GDPR) and Proposed AI Act

On March 13, 2024, the European Parliament approved the Artificial Intelligence Act (AI Act), a major piece of legislation aimed at shaping AI use within the European Union. Passed with 523 votes in favor, 46 against, and 49 abstentions, the Act establishes a detailed regulatory structure to protect public interests, fundamental rights, and environmental standards in response to high-risk AI applications. The Act is designed to guide AI's growth in a responsible manner, placing Europe at the forefront of global AI policy (European Parliament, 2024).

The Act prohibits AI applications considered particularly harmful to citizens' rights and privacy. Prohibited uses include biometric categorization based on sensitive characteristics, large-scale facial recognition through untargeted data scraping, and emotion recognition in work or educational settings. Social scoring, profiling-based predictive policing, and AI designed to exploit user vulnerabilities are also banned. While real-time biometric identification (RBI) by law enforcement is generally restricted, it may be used in cases such as locating missing persons or preventing terrorism, provided it is under strict limitations with judicial oversight.

High-risk AI systems—those impacting critical areas like healthcare, infrastructure, and education—must follow clear obligations for risk management, transparency, and human

oversight. Citizens affected by such systems have the right to file complaints and receive explanations, reflecting the EU's emphasis on transparency and individual rights.

Transparency is a priority, especially for general-purpose AI (GPAI) models, which are required to disclose content used in their training and adhere to copyright standards. Manipulated content like deepfakes must be clearly labeled, and systems that could have extensive social impacts face more stringent monitoring, including regular assessments and incident reporting.

To support innovation, particularly among small businesses, the Act introduces "regulatory sandboxes" that allow AI testing in controlled, real-world environments, with special provisions for small and medium-sized enterprises (SMEs). During the plenary session, co-rapporteur Brando Benifei described the legislation as the world's first binding AI law, designed to minimize risks while keeping European values central to AI development. Benifei highlighted the upcoming AI Office, which will help companies transition to compliance, while co-rapporteur Dragos Tudorache noted that the AI Act sets the foundation for a new governance approach focused on technology.

After a final legal review and Council approval, the AI Act will take effect 20 days post-publication, with phased deadlines: prohibited practices within six months, governance standards within 12 months, and high-risk system requirements within 36 months. The Act aligns with citizens' recommendations from the Conference on the Future of Europe, responding to calls for responsible AI that enhances transparency, safety, and European competitiveness.

India: AI Strategy and Regulatory Framework

As of March 2024, India has yet to establish dedicated AI legislation. Instead, AI oversight depends on advisories, guidelines, and existing IT regulations. On March 1, 2024, the government issued guidance requiring platforms to obtain clearance from the Ministry of Electronics and Information Technology (MeitY) before deploying untested AI systems or large language models (LLMs) to the public. This directive mandates that intermediaries avoid promoting bias, maintain electoral integrity, and label content produced by AI. Minister Rajeev Chandrasekhar later clarified that this requirement primarily targets larger platforms, excluding startups. Following public feedback, the government revised the directive, removing the need for action reports while preserving obligations for user alerts about unverified AI models and labeling of deepfake content (Dey & Cyrill, 2024).

India's AI regulatory environment draws on several foundational initiatives. The National Artificial Intelligence Strategy, branded as #AIFORALL and launched by National Institution for Transforming India (NITI Aayog) in 2018, highlights healthcare, agriculture, education, and transportation as focal areas for AI-driven advancements (Bhalla et al., 2023). This strategy led to data quality improvements and preliminary structures for cybersecurity and data protection. Building on these, NITI Aayog introduced Principles for Responsible AI in 2021, which lay out

ethical guidelines across seven key areas: safety, transparency, accountability, inclusivity, privacy, equality, and reinforcing positive societal values.

Further policy expansion came with the Digital Personal Data Protection (DPDP) Act in August 2023, targeting digital data security and safeguarding personal information. Alongside, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, updated in April 2023, form a regulatory foundation for supervising social media, digital media intermediaries, and other related platforms, strengthening India's responsible AI usage framework.

India has also committed to significant financial backing for AI, recently approving INR 103 billion (around USD 1.25 billion) over five years to build AI infrastructure, support LLMs, and promote AI innovation among startups. A National Data Management Office (NDMO) is set to collaborate across government departments, focusing on improving the quality and accessibility of data for AI applications within the public sector.

To address potential risks from emerging AI technologies, India's Ministry of Electronics and Information Technology and the Bureau of Indian Standards are working to establish national AI standards. These standards aim to create comprehensive guidelines on AI ethics and safety. Although India lacks direct laws targeting deepfakes—AI-manipulated content that can impact personal and public trust—existing provisions within the Information Technology Act, 2000, and sections of the Indian Penal Code (IPC) provide remedies. For example, Sections 66E and 66D of the IT Act address privacy invasions and misuse of digital resources, while IPC Sections 509 and 499 cover related offenses like defamation and misinformation.

On March 15, 2024, MeitY released an updated advisory reinforcing the need for intermediaries to manage AI-generated and modified content responsibly. This advisory calls for platforms to prevent bias, maintain election integrity, and label AI-created deepfake content with permanent metadata. Non-compliance with the IT Act, 2000, or its associated IT Rules may result in legal consequences for platforms, intermediaries, and their users.

Japan: AI Strategy and Policy Framework

As of July 2024, Japan's approach to AI regulation relies on "soft law," prioritizing voluntary adherence to guidelines over binding regulations. While there is currently no dedicated AI legislation, recent frameworks have been introduced to steer AI developers, providers, and users toward ethical and responsible practices. Notably, on April 19, 2024, Japan released *AI* Guidelines for Business Version 1.0, combining previous guidelines into a single set of standards that advocates for an "agile governance" model. This model calls on stakeholders to engage in continuous analysis, objective-setting, and evaluation to manage AI systems responsibly (Kamiya & Keate, 2024).

Japan has also issued the Hiroshima International Guiding Principles for Organizations Developing Advanced AI Systems, which provides international guidance for safe and

trustworthy AI. To explore future AI regulations, Japan's AI Strategy Council proposed draft recommendations on May 22, 2024, assessing regulatory options for high-risk AI technologies. A government working group also proposed the Basic Act on the Advancement of Responsible AI, which, if passed, would establish direct regulatory controls on certain generative AI models. This act would introduce specific requirements for developers, such as rigorous evaluations, operational standards, and regular reporting, marking a shift toward "hard law" for high-impact AI systems. Enforcement can involve government monitoring and penalties for non-compliance, contrasting with Japan's previous reliance on voluntary compliance (Kamiya & Keate, 2024).

While Japan lacks specific AI laws, other legislative measures influence AI applications indirectly. The Digital Platform Transparency Act mandates transparency in large online platforms' transactions, and the Financial Instruments and Exchange Act imposes risk management standards on high-speed trading entities. Additionally, the Civil Code allows for defamation claims involving AI, and the Copyright Act and the Act on the Protection of Personal Information address unauthorized use of personal and copyrighted content. On May 9, 2024, the Information Distribution Providers Act—an update to Japan's Providers' Liability Limitation Act—was passed, aiming to streamline requests for removing harmful content online. While not AI-specific, this law is referenced in policy discussions as a tool for addressing misinformation and AI-generated content risks. Japan's Criminal Code also applies to AI-related offenses, such as defamation, business obstruction, and unauthorized control of others' devices (Kamiya & Keate, 2024).

The new guidelines emphasize core principles for AI that are human-centric, stressing respect for human rights, safety, and fairness, alongside privacy and data security. The guidelines also highlight the importance of transparency, accountability, fair competition, and ongoing innovation in AI development. Japan's Ministry of Economy, Trade, and Industry, the Ministry of Internal Affairs and Communications, and the Agency for Cultural Affairs oversee these AI governance measures, playing an advisory role in implementing these guidelines.

If enact, the Basic Act on the Advancement of Responsible AI can significantly expand Japan's AI oversight, with potential penalties for non-compliance, marking a shift from flexible guidance to structured enforcement. This evolution in policy reflects Japan's intent to balance ethical safeguards with technological advancement as AI's societal impact grows.

OECD: Principles on Artificial Intelligence

The Recommendation of the Council on Artificial Intelligence, established by the Organisation for Economic Co-operation and Development (OECD) and adopted by 46 countries as of July 2021, provides a foundational framework for AI governance. Known as the "Principles," these guidelines represent the first intergovernmental AI standard, later serving as a model for the G20's AI Principles. The OECD's framework promotes trustworthy AI by urging Adherents to embed human-centered values, transparency, and accountability in their national AI policies and international AI collaborations. While not legally binding, these principles reflect a significant

commitment to ethical AI practices across member and non-member states (Hickman, Zaidi, & Mair, 2024).

The OECD's Recommendation includes five specific policy areas to support ethical AI practices: (1) increased investment in AI research and development, (2) access to a secure digital ecosystem, (3) a flexible policy environment that supports AI throughout its lifecycle, (4) workforce preparedness for AI's societal and economic impacts, and (5) collaboration among international partners. The framework enables each country, referred to as Adherents, to adapt these standards to its own legal and regulatory systems. The Recommendation has influenced initiatives such as the G7's Hiroshima AI Process and the International Guiding Principles on AI, highlighting the OECD's role in shaping unified global AI standards. Although the OECD's guidelines are non-binding, Adherents are expected to actively implement them. The AI Policy Observatory, an OECD platform, monitors member countries' progress, providing a dynamic database of AI policies, metrics, and strategies. This resource enables governments to exchange knowledge, enhance policy clarity, and align with international standards. The observatory supports ongoing policy development, encouraging Adherents to adopt effective measures for ethical AI governance (Hickman, Zaidi, & Mair, 2024).

To address evolving technologies, including generative AI, the OECD updated its AI system definition in November 2023. The new definition covers the complete AI lifecycle, from design and data processing through to deployment and monitoring, allowing for detailed oversight across AI's operational phases. Additionally, the recommendation defines key terms such as "AI actors"—individuals and organizations involved in the lifecycle—and "AI knowledge," which refers to essential skills, resources, and processes necessary for effective AI governance. The recommendation's reach is comprehensive, as it applies to various sectors without specific industry limitations, providing a consistent framework for AI use across different fields. While enforcement mechanisms are absent, the OECD expects Adherents to integrate these principles and recommendations into their national standards, creating an internationally aligned approach to ethical AI (Hickman, Zaidi, & Mair, 2024).

Russia: AI Development and Regulation

On March 1, 2024, Russian President Vladimir Putin signed a detailed 40-page decree to revise Russia's national AI strategy, extending its framework through 2030. This Presidential Decree introduces updates to the National Strategy for Artificial Intelligence Development in the Russian Federation, positioning AI as a pivotal element in advancing the nation's economic and technological landscape. The updated strategy includes requirements for federal agencies, encouraging the adoption of AI in various sectors and regions of Russia (Werner, 2024).

One of the decree's main provisions sets a deadline of July 1, 2024, by which AI integration must be embedded within the national data economy program. Federal bodies are directed to align their strategic documents and sectoral planning initiatives with this AI framework. Local

and regional governments are encouraged to incorporate AI into digital transformation plans, and state-owned enterprises are urged to prioritize AI development in their operational strategies.

The updated AI strategy emphasizes core principles such as safeguarding human rights, enhancing security, achieving technological independence, and promoting fair competition. The decree establishes new objectives that focus on expanding AI infrastructure access, supporting AI developers, advancing AI-related research and workforce training, encouraging adoption of AI across various sectors, ensuring safe AI practices in high-risk scenarios, and constructing a comprehensive legal foundation to support these efforts.

The decree also introduces new performance indicators, including metrics for supercomputing capacity, the impact of AI on GDP, growth in the AI services market, the volume of research publications, workforce training outcomes, public confidence in AI, and the rate of enterprise adoption.

Russia's strategy reflects both advancements and persistent challenges in the field of AI. It highlights achievements in AI education, startup support, and applications in areas like healthcare, while also identifying issues such as limited computational resources, reliance on imported technology, talent shortages, slow adoption by government agencies, legal obstacles, data quality limitations, and cybersecurity risks.

In the governance section, updates focus on ethical AI standards, safety protocols, and balancing public and private sector priorities. To increase access to AI infrastructure, the decree promotes initiatives like cloud computing, providing discounted computing resources for students, startups, and researchers, boosting local electronics production, and enhancing data availability and quality.

The strategy sets out significant support for AI developers, including grants, investment incentives, entrepreneurial skill development, commercialization of research, and open-source resources. Research efforts will be strengthened with funding for cutting-edge projects, interdisciplinary and core AI research, corporate R&D support, talent attraction, and model development. Publication metrics and collaborative research initiatives will be used to assess the quality of AI advancements.

United Regulations in Asia: ASEAN

In February 2024, the Association of Southeast Asian Nations (ASEAN), which includes Brunei, Cambodia, Indonesia, Lao PDR, Malaysia, Myanmar, the Philippines, Singapore, Thailand, and Vietnam, released the ASEAN Guide to AI Ethics and Governance. This guide is intended to help organizations across these countries responsibly design, develop, and deploy AI, while also providing a framework for governments to shape AI policies. ASEAN's approach is flexible and non-binding, encouraging ethical AI practices without enforcing strict requirements (Lauw, Ching, & Cheng, 2024).

The ASEAN Guide focuses on traditional AI technologies and does not yet address generative AI. It shares similarities with Singapore's Model AI Governance Framework and includes recommendations at both the national and ASEAN-wide levels. Instead of defining high-risk AI categories, as seen in the EU's AI Act, it advises companies to respect cultural differences among ASEAN member states, allowing each country to customize AI ethics for its own context.

Generally, AI regulation across Asia remains light and largely voluntary, as many governments seek to promote AI industry growth. However, countries like Vietnam and South Korea are exploring a shift towards more structured regulation. This regional trend may strengthen as countries observe and evaluate the EU's AI Act, potentially influencing future regulatory decisions.

Discussion

The development and deployment of AI products present numerous parameters and conflicting issues that are challenging to address. These include the dynamic structure of AI systems, the fast-growing industry and competition, and the AI's black box dilemma, which necessitates transparency standards. Additionally, the need for explainability tailored to diverse stakeholder needs, cultural differences, and the creation of universally applicable rules further complicate the landscape. High-quality documentation is essential but balancing ethical and financial priorities remains a complex task. Context-specific standards and legislation must be considered, along with enforcement mechanisms that may be compulsory or voluntary. Practical considerations for organizations and developers, advancing responsible AI through diverse and ethical development teams, and the secret agendas of competing actors also play significant roles. Finally, the monopoly of technology and universal needs must be addressed to ensure equitable and fair AI development.

Opinions of Global Leaders at World Economic Forum 2024 in Davos

At the 2024 World Economic Forum in Davos, global leaders and experts engaged in robust discussions about the complexities of AI regulation, debating whether oversight should focus on the technology itself or its applications and impacts. Andrew Ng, founder of DeepLearning.AI, cautioned against directly regulating AI development, warning that doing so could stifle innovation and disproportionately benefit large tech companies. Ng (2023) argued that excessive constraints, particularly on open-source software, might impede the societal benefits of AI. Instead, he advocated for application-specific regulation, a view echoed by Khalfan Belhoul, CEO of the Dubai Future Foundation, who noted the abstract nature of AI makes it difficult to regulate as a whole. Belhoul proposed governing AI's sectoral impacts through tailored, case-by-case regulation, which he argued would be more practical and effective (Ruggeri, 2024).

Conversely, Arati Prabhakar, Director of the U.S. Office of Science and Technology Policy, stressed the importance of proactive oversight of AI as a transformative technology with the potential for both societal benefit and harm. She emphasized the U.S. government's

responsibility to ensure AI advances the public good while mitigating risks. In the healthcare sector, for instance, existing regulations that enforce a "duty of care" for physicians could be extended to AI applications. Wendell Wallach, a scholar at the Carnegie Council for Ethics in International Affairs, observed that many AI applications already fall under existing frameworks such as healthcare and consumer protection laws. However, Wallach argued for additional safeguards, including compliance testing and quality assessments, to address AI's unique risks (Ruggeri, 2024).

Challenges, Tradeoffs, and Opportunities

Although the awareness of the ethical challenges posed by the emergence of generative AI has grown in the wake of the arrival of tools like ChatGPT, there is a gap between this awareness and actual action around developing comprehensive regulatory structures in this area. AI, much like other technologies emerging in recent years such as nuclear energy and biotechnology, carries immense potential for positive outcomes if used responsibly, but could have tremendously detrimental impacts if utilized by humans with nefarious or misguided intent (Wang et al., 2018). AI, as a generic technology in itself, is neither dangerous nor benign. Its impact is dependent on the responsible usage and thoughtful application. Ethical integrity and transparency are essential in guiding AI toward beneficial purposes and calling for engaged, conscientious participation from a broad cross-section of society.

AI's Black Box Dilemma and Transparency Standards

AI's "black box" dilemma, where no transparency exists about how these tools operate, diminishes public trust and developer accountability (Castelvecchi, 2016). Many standards and policies have been proposed around the world in recent years to combat the issues observed from this dilemma. Among the currently proposed AI transparency standards, several recurring themes emerge, which provide valuable insight into the global landscape of discourse about transparency oversight in AI. Notably, all standards frameworks investigated in this paper identify transparency as a cornerstone for providing public trust and developer accountability in AI systems, situating policy at the nexus of the successful diffusion of this technology. While debates persist about the optimal level of oversight needed, there is a broad consensus that some form of transparency is requisite.

Explainability and Stakeholder Needs

The proposed AI transparency standards and frameworks emphasize the principle of explainability: that those who work with AI systems should have access to information about how the systems arrive at decisions. These explanations should be tailored to the diverse information needs of various stakeholders (Ridley, 2022). The technical auditor of these systems, such as a government oversight agency under the proposed U.S. Research, Innovation, and Accountability Act, requires comprehensive and detailed information about the AI systems being monitored. This information is essential for evaluating compliance with established standards

and assessing the safety and reliability of the systems. However, end-users of these systems, who may lack technical expertise, require a simplified, jargon-free explanation of how these systems function. All stakeholders require information about what data was used to train the system and the processes used in the training (Daneshjou et al., 2021). This is necessary for users not only to make an informed decision about whether to adopt a specific AI technology, but also to determine whether they want their data to be used for training purposes. In contexts such as business and healthcare, managers may require precise and context-specific information to determine whether the tool is appropriate for their specific applications and complies with legal and ethical standards.

The Importance of High-Quality Documentation

Another critical aspect emphasized in numerous proposed and active AI transparency standards is the significance of high-quality and consistent documentation. This requirement is most prominently demonstrated in the CLeAR framework but is also reflected in several other standards. Documentation about AI systems must be both comprehensible and actionable (Winecoff & Bogen, 2024). These proposed standards advocate for regular updates of documentation as an AI system advances through all stages of its life cycle, to ensure continuous transparency. This continuous and accessible documentation ensures that stakeholders always have access to the most current information about the AI system, which is vital in ensuring that these groups can make sound decisions and maintain trust in AI technologies.

Balancing Ethical and Financial Priorities

Ultimately, these mandates for enhanced AI transparency must be balanced with a range of other forces and ethical considerations. Developers want to ensure their intellectual property rights, as AI models represent great value. The right to data privacy and robust security measures also requires that some data and processes behind AI models be kept private to protect users and developers alike (Kazim & Koshiyama, 2021). Striking the proper balance between competing ethical and financial priorities with AI development is a complex, and multi-faceted challenge (Borenstein & Howard, 2021). Achieving greater transparency will result in trade-offs in the other areas. The complexity of this issue is heightened by the varying priorities of different stakeholders. Developers may prioritize protecting their intellectual property. Academics and regulators may emphasize and advocate for the highest level of transparency about AI models. End-users are more likely to simply seek accessible information on the models that is relevant to their levels of knowledge and usage of the tools. Balancing these perspectives requires collaboration and ongoing dialogue among all of these groups.

Context-Specific Standards and Legislation

Additionally, developers and regulators alike must consider the context in which AI systems will be employed. Certain contexts in the Global North may require different standards than in the Global South, and vice versa (Mannuru et al., 2023). Organizational contexts like higher

education, medicine, and finance may require stricter standards than other contexts (Lund et al., 2023). These obligations are exhibited in several of the proposed standards and legislation described in this paper. For example, the HTI-1 Final Rule provides standards that are more stringent than some others, but are important for the unique situation of medicine, where patient data is strictly controlled by law in many countries.

Enforcement Mechanisms: Compulsory vs. Voluntary Compliance

Notably, the frameworks discussed in this paper vary in their mechanisms of enforcement and their compulsory versus voluntary nature. For instance, the Research, Innovation, and Accountability Act proposed by the United States' Congress would make developers legally responsible for transparency, using fines and closures to enforce their standards. While this approach may ensure compliance, it could also create rigidity that may stifle further innovation. Conversely, voluntary compliance with measures like the IEEE and CLeAR frameworks could allow flexibility to adopt some aspects while modifying other standards that may slow innovation. However, the ability to modify or limit compliance with certain standards could lead to ethical challenges and a lack of an enforcement mechanism could mean that some developers would ignore the framework altogether. If the objective of discussing transparency frameworks in the first place is to ensure accountability, then having completely voluntary standards may be unsuitable. Alternatively, having standards akin to the Web Content Accessibility Guidelines (WCAG) (Caldwell et al., 2008), with multiple and minimal levels of compliance, may be more amenable.

Practical Considerations for Organizations and Developers

From a practical standpoint, organizations that utilize generative artificial intelligence tools may place pressure on developers to ensure transparency by including this as a criterion used in requests for proposals for new systems, as they might do with the WCAG for accessibility. Developers are likely to be motivated by financial considerations, so requirements to adhere to a specific set of standards will help make those standards commonplace within the industry. Organizations involved in developing generative AI tools may want to preemptively steer the development of their models and documentation to conform to the popular standards and frameworks to reduce workload when these regulatory structures inevitably become common practice. Conformance to new standards guiding an industry through legal compliance review and product redesign efforts are considered "soft costs" that can take a substantial toll on developers (Zahirah et al., 2013). Preemptive changes may help to spread these soft costs over several years.

Advancing Responsible AI through Diverse and Ethical Development Teams

A crucial factor in advancing responsible AI with a high level of transparency lies in the composition and preparation of the teams developing these technologies. Developers and stakeholders benefit from training that not only enhances technical expertise but also builds

cultural awareness and ethical insight. Such education encourages diversity within development teams, which reduces unintentional biases and facilitates more inclusive designs. The inclusion of policymakers and ethics experts from fields like medical science and systems on development teams would help mitigate the hidden biases that currently plague development in AI.

Conclusion

The rapid evolution of AI has left many policymakers unprepared for the ethical pitfalls associated with these technologies. A global push for increased transparency in AI models reflects a shared recognition of its critical role in building public trust and holding developers accountable. Despite growing awareness of AI's ethical challenges, there remains a gap between this awareness and the development of comprehensive regulatory structures. AI's potential for both positive and negative impacts emphasizes the need for responsible usage and thoughtful application.

Transparency is crucial in addressing the "black box" dilemma, which diminishes public trust and developer accountability. Proposed standards and frameworks consistently identify transparency as essential for public trust and accountability, emphasizing explainability tailored to diverse stakeholders and high-quality, consistent documentation. Balancing transparency with intellectual property rights, data privacy, and security is a complex challenge. Different stakeholders, including developers, academics, regulators, and end-users, have varying priorities, necessitating collaboration and ongoing dialogue. Context-specific standards are also important, as different regions and sectors may require different levels of oversight.

Enforcement mechanisms vary, with some frameworks advocating for compulsory compliance and others for voluntary adherence. Practical considerations for organizations include incorporating transparency criteria in requests for proposals and preemptively aligning with popular standards to reduce future regulatory burdens. Training development teams in technical expertise, cultural awareness, and ethical insight is also crucial for advancing responsible AI.

The challenges remain in balancing transparency with privacy, security, and intellectual property rights while significant strides have been made in developing transparency standards, frameworks, and legislation. Future work should focus on refining these frameworks to ensure they are adaptable to various contexts and stakeholder needs. Continued global collaboration and dialogue will be essential in developing effective regulatory structures that promote transparency and accountability in AI, ultimately advancing a future where technology contributes positively to society.

References

Artificial Intelligence Research, Innovation, and Accountability Act of 2023, S. 3312 (2024).

Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2023). The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics, 3*(2), 389-406.

Australian Government. (2021). *Artificial Intelligence Ethics Framework*. Retrieved from https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology, 159*, article 107197.

Bareis, J., & Katzenbach, C. (2022). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values, 47*(5), 855-881.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021*, 610–623.

Bhalla, N., Brooks, L., & Leach, T. (2023). Ensuring a 'Responsible' AI future in India: RRI as an approach for identifying the ethical challenges from an Indian perspective. *AI and Ethics, 4*, 1409-1422.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 149-159). Association for Computing Machinery.

Booker, C. (2023). U.S. *Senate Introduces the Algorithmic Accountability Act*. https://www.booker.senate.gov

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics, 1*, 61-65.

Caldwell, B., Cooper, M., Reid, L. G., Vanderheiden, G., Chisholm, W., Slatin, J., & White, J. (2008). Web content accessibility guidelines (WCAG) 2.0. WWW Consortium (W3C), 290(1-34), 5-12.

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News, 538*(7623), 20-23.

Chmielinski, K., Newman, S., Kranzinger, C. N., Hind, M., Vaughan, J. W., Mitchell, M., ... & Chang, A. (2024). *The CLeAR Documentation Framework for AI Transparency.* Shorenstein Center on Media, Politics and Public Policy. https://shorensteincenter.org/clear-documentation-framework-AI-transparency-recommendations-practitioners-context-policymakers/

Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.

Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatology*, *157*(11), 1362-1369.

Dey, A., & Cyrill, M. (2024, March 27). *India's regulation of AI and large language models*. India Briefing. https://www.india-briefing.com/news/india-regulation-of-ai-and-large-language-models-31680.html/

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019, March). Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 275-285). Association for Computing Machinery.

EU Commission. (2021). Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence. *Brussels*, *21*, 2021.

European Parliament. (2024, March 13). *Artificial Intelligence Act: MEPs adopt landmark law*. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

Everson, J., Smith, J., Marchesini, K., & Tripathi, M. (2024). A regulation to promote repsonsible AI in health care. *Health Affairs*. https://doi.org/10.1377/forefront.20240223.953299

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333-3361.

Fernandez-Quilez, A. (2023). Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. *AI and Ethics*, *3*(1), 257-265.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In Meyers, G., & Beiker, S., *Road vehicle automation* (pp. 93-102). Springer.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Government of Canada. (2021). *Directive on Automated Decision-Making*. Retrieved from https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Hickman, T., Zaidi, Z., & Mair, D. (2024, May 13). *AI Watch: Global regulatory tracker - OECD*. White & Case. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-oecd

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, *1*(1), 41-47.

Hind, M., Houde, S., Martino, J., Mojsilovic, A., Piorkowski, D., Richards, J., & Mojsilović, A. (2020). Experiences with Improving the Transparency of AI Models and Services. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-8). Association for Computing Machinery.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. arXiv preprint arXiv:1805.03677.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575). Association for Computing Machinery.

IEEE Standards Association. (2022). IEEE standard for transparency of autonomous systems. *IEEE Std*, 7001-2021, pp.1-54.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Kamiya, M., & Keate, J. (2024, July 1). *AI Watch: Global regulatory tracker - Japan*. White & Case. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-japan

Kazim, E., & Koshiyama, A. (2021). The interrelation between data and AI ethics in the context of impact assessments. *AI and Ethics, 1*, 219-225.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review, 9*(2), 1-16.

Lauer, D. (2020). You cannot have AI ethics without ethics. *AI and Ethics, 1*, 21-25.

Lauw, N., Ching, P. F., & Cheng, A. (2024, August 6). *Part 4 – AI Regulation in Asia*. RPC. https://www.rpclegal.com/thinking/artificial-intelligence/ai-guide/part-4-ai-regulation-in-asia

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570-581.

Luong, N. (2024, August 29). *China's AI governance: Engaging the global South*. National Bureau of Asian Research. https://www.nbr.org/publication/chinas-ai-governance-engaging-the-global-south/

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Association for Computing Machinery.

Mannuru, N. R., Shahriar, S., Teel, Z. A., Wang, T., Lund, B. D., Tijani, S., ... & Vaidya, P. (2023). Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. Information Development, early view. https://doi.org/10.1177/02666669231200628.

Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence, 5*, article 100152.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). Association for Computing Machinery.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). 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*, *26*(4), 2141-2168.

National Artificial Intelligence Advisory Committee. (2024). *Towards standards for data transparency for AI models*. https://ai.gov/wp-content/uploads/2024/06/PROCEEDINGS_Towards-Standards-for-Data-Transparency-for-AI-Models.pdf

Ng, A. (2023) *Written Statement of Andrew Ng Before the U.S. Senate AI Insight Forum, December 11, 2023.* https://aifund.ai/insights-written-statement-of-andrew-ng-before-the-u-s-senate-ai-insight-forum/

Okolo, C. T. (2024, March 15). *Reforming data regulation to advance AI governance in Africa*. Brookings. https://www.brookings.edu/articles/reforming-data-regulation-to-advance-ai-governance-in-africa

Pagallo, U. (2017). The legal challenges of big data: putting secondary rules first in the field of EU data protection. *European Data Protection Law Review*, *3*, 36.

Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1776-1826). Association for Computing Machinery.

Quinn, M., Piper, B., Bliss, J. P., & Keever, D. (2020, December). Recommended methods for using the 2020 NIST principles for ai explainability. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 2034-2037). IEEE.

Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics, 3*(3), 735-744.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). Association for Computing Machinery.

Richards, J., Piorkowski, D., Hind, M., Houde, S., & Mojsilović, A. (2020). A Methodology for Creating AI FactSheets. arXiv preprint arXiv:2006.13796.

Ridley, M. (2022). Explainable Artificial Intelligence (XAI): Adoption and Advocacy. *Information Technology and Libraries, 41*(2), 1-17.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Ruggeri, A. (2024, January 19). *Davos 2024: Can – and should – leaders aim to regulate AI directly?* World Economic Forum, https://www.bbc.com/worklife/article/20240118-davos-2024-can-and-should-leaders-aim-to-regulate-ai-directly

Schiff, D. (2022). Education for AI, not AI for education: The role of education and ethics in national AI policy strategies. *International Journal of Artificial Intelligence in Education, 32*, 527-563.

Shin, D. (2019). Toward fair, accountable, and transparent algorithms: Case studies on algorithm initiatives in Korea and China. *Journal of the European Institute for Communication and Culture, 26*(3), 274-290.

Srinivasan, R., & Ghosh, D. (2023). A new social contract for technology. *Policy & Internet, 15*(1), 117-132.

Stoyanovich, J., & Howe, B. (2019). Nutritional Labels for Data and Models. *IEEE Data Engineering Bulletin, 42*(3), 13-23.

Swaminathan, N., & Danks, D. (2024). Application of the NIST AI Risk Management Framework to Surveillance Technology. arXiv preprint arXiv:2403.15646.

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science, 29*(3), 230-241.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: key problems and solutions. *AI and Society, 37*, 215-230.

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), 1607-1622.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76-99.

Wang, Q., Li, R., & He, G. (2018). Research status of nuclear power: A review. *Renewable and Sustainable Energy Reviews, 90*, 90-96.

Werner, J. (2024, March 1). *Russia updates national AI strategy*. Babl AI. https://babl.ai/russia-updates-national-ai-strategy/

Winecoff, A. A., & Bogen, M. (2024). Improving governance outcomes through AI documentation: Bridging theory and practice. arXiv preprint arXiv:2409.08960.

Winfield, A. F., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., ... & Watson, E. (2021). IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, *8*, 665729.

Wulf, A. J., & Seizov, O. (2020). Artificial intelligence and transparency: A blueprint for improving the regulation of AI applications in the EU. *European Business Law Review, 31*(4), 611-640.

Wyden, R. (2023). *Algorithmic Accountability Act Summary*. Retrieved from https://www.wyden.senate.gov/imo/media/doc/algorithmic_accountability_act_of_2023_summary.pdf

Yekaterina, K. (2024). Challenges and opportunities for AI in healthcare. *International Journal of Law and Policy, 2*(7), 11-15.