A Scoping Review of Transparency and Explainability in AI Ethics Guidelines

Kerrie Hooper, Stephanie Lunn
Florida International University
Miami, FL
khooper@fiu.edu, sjlunn@fiu.edu

Abstract

Transparency and explainability are crucial tenets of ethical Artificial Intelligence (AI) and are often classified as technical components of AI Ethics. Many countries and international governing bodies have developed AI guidelines and principles that are made public for respective civilians with a diverse range of expertise and knowledge. This short paper compares how explainability and transparency are presented and discussed in AI ethics guidelines developed by the top ten countries leading in AI research and development according to the AI Global Index Report in 2023 and leading efforts from governing bodies such as the EU. Methodologically, this paper presents a thematic analysis focusing on the presence and acknowledgment of various dimensions of explainability and transparency, as well as the level of detail and examples in the guidelines. The aim is to uncover how various aspects of AI ethics are presented in global guideline documents and highlight areas in which the documents converge and discuss implications for AI stakeholders.

Introduction

In the context of ethical Artificial Intelligence (AI), *transparency* suggests methods to support a clear understanding of how the AI system works, makes decisions and handles data (Lawton, 2023). *Explainability* entails providing sound reasons for the decisions made by the AI system (Grennan et al., 2022). Both neighboring concepts are considered technical dimensions of ethical AI, where transparency is more of an umbrella term supported by explainability. However, while explainability is often at the algorithmic level, transparency is broader and refers to the AI system at the larger level, albeit at the organizational or national level (Woudstra, 2020). A less ambiguous term would be algorithm transparency. Still, a multidisciplinary approach is needed to address transparency since explainable AI research typically does not build on frameworks from the social sciences or humanities, and more work can be done to expand our understanding of transparency in AI (Larsson & Heintz, 2020). In addition, a multidisciplinary approach would contribute to notions of trust and reliability to end users and society (Mohseni et al., 2021). Many nations worldwide have adopted AI for technological efficiency and economic development. This is evident as many countries have published their national AI strategies. The Organization for Economic Cooperation and Development - OECD.AI (2021) created a database of policy initiatives from 69 countries and the European Union (EU), which published national AI strategies and policies, with most of those policies coming from governmental entities. Additionally, now that AI ethics is at the forefront, there have been many added initiatives surrounding the responsible development and deployment of AI at various institutional, national, and global levels. For instance, the United Nations Educational, Scientific and Culture Organization (UNESCO) published a report drafting a recommendation on the ethics of AI (Ad Hoc Expert Group, 2020). Therefore, international efforts exist, and an agreement exists on the necessity of these initiatives as discussions on ethical AI begin to permeate globally. Many nations work together to adhere to guidelines and continue to develop the AI ethics landscape, forming institutions such as the Global Partnership on Artificial Intelligence (GPAI) and many others (Schmitt, 2022). Even though these global partnerships are tremendously important, individual countries have been showcasing efforts and contributions toward the ethical AI movement through guidelines, standards, or policies. These documents are often intended to guide and support AI stakeholders in developing ethical AI. Deshpande and Sharp (2022) classified AI stakeholders into three levels: a) Individual stakeholders, such as users, developers, engineers, researchers, AI experts, and non-experts; b) Organizational stakeholders, like technology companies, professional organizations such as the Institute of Electrical and Electronics Engineers (IEEE); and c) National/International stakeholders who create laws and regulations such as nation states and regulative agencies like the United Nations International Children's Emergency Fund (UNICEF).

It is vital to note that a range of AI stakeholders exists, as this helps in understanding the broad lens of people for whom these policies and guidelines are designed. This paper focuses on the Individual and Organizational stakeholders zoning in on the developers, engineers, and various end users as these groups undergo technical workloads and are the recipients of the technological impact of AI while having to consider multiple guidelines they are expected to abide uphold. Accordingly, we sought to explore a subset of global policy documents on transparency and explainability, seeking to answer the following research question (RQ):

RQ: In what ways did countries leading in AI ethics discuss transparency and explainability in their AI ethics guidelines?

To answer this question, we conducted an exploratory study of guidelines from different countries worldwide. We employed thematic analysis to understand how a subset known to be engaged in AI ethics addressed these topics. While we acknowledge there may be many approaches to understanding the current state of affairs, this work is intended as a first step in examining AI ethics as it relates to regulations and recommendations for different stakeholders from a global context. Findings from investigations across countries may serve as insightful to broadening the structure and scope of AI ethics guidelines and standards.

Related Work

A fundamental way in which studies described transparency and explainability is that they are in the form of a hierarchy where explainability can be interpreted as a method of transparency (Ehsan et al., 2021, p. 2; Mc Dermid et al., 2021, p. 5). Balasubramaniam et al. (2022) conducted a study to determine what AI ethics guidelines are developed for transparency and explainability and how those concepts contribute to quality requirements. The authors assessed 16 AI ethics guidelines created at the organizational level and found that transparency and explainability were vital quality requirements that can achieve the goal of trustworthiness. Additionally, the authors noted that it can be challenging to distinguish transparency and explainability based on their analysis of ethical guidelines.

Ehsan et al. (2021) argued for the inclusion of "socio" in the sociotechnical context of explainable AI (XAI). They introduced the concept of socio-transparency (ST) to incorporate a socio-organizational context to enable holistic XAI systems. The authors also studied the potential effect of ST by asking systems-related questions by utilizing the 4Ws (who, what, why, and when) through a series of interviews with AI stakeholders.

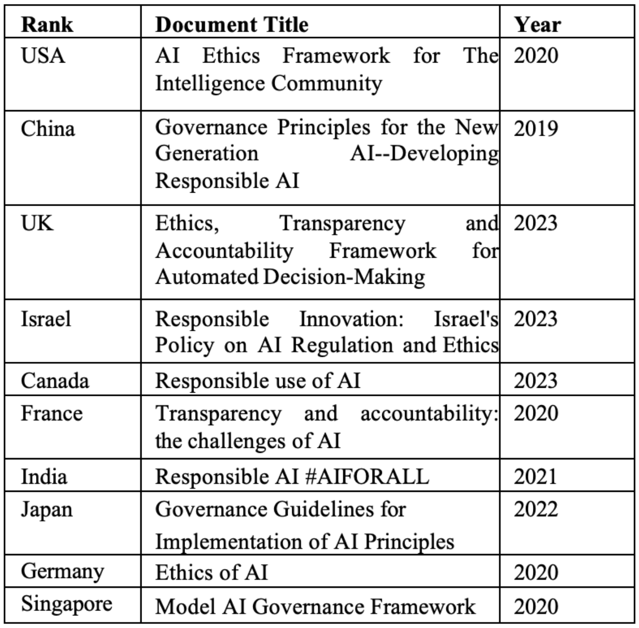
Mc Dermid et al. (2021) presented an overview of XAI methods and discussed why the stakeholders might need them. They noted that stakeholders might need XAI to clarify results, check compliance and confidence, be aware of consent and control, and challenge the AI results. The authors then explained various XAI methods, such as feature importance and example-based methods. An important point they highlighted is that needsstakeholder needs often contrast with the capabilities of the XAI methods.

Expanding ethical insights on a multinational level is crucial to developing global trust in AI. Yet, at the design, policy, and implementation levels, it is well established that achieving ethical AI is no simple task. The Global AI Index Report 2023 (Cesareo & White, 2023) recognized 62 countries leading in AI development and research. These countries have also produced various responsible AI guidelines for AI stakeholders. Additionally, assessing these guidelines aids in bringing together global perspectives to expand our understanding of and implement ethical AI. This study is unique because it focuses on AI guidelines from top AI-achieving countries around the globe regarding research and development.

Method

AI ethics guidelines were gathered from the top ten ranking countries in the Global AI Index Report to explore how transparency and explainability are considered in international policy documents. To find these publicly available guidelines, the search terms were in the manners of "AI Ethics Guidelines {country_name}" and "Responsible AI Guidelines {country_name}." The first guideline document that appeared in the search using Google's search algorithm, which fit the following inclusion criteria, was used in this study. For government principles and guidelines to be included in this study, they had to be centered around AI ethics or responsible AI specifically and not be general to algorithms or Information and Communications Technology (ICT) strategies. Additionally, the guidelines had to be derived from a source supported by the respective government, preferably from sources ending in ".gov" or strongly references government sources. The goal was to analyze one policy per country. Essentially, the documents found were thoroughly analyzed to ensure they came from a reputable government source.

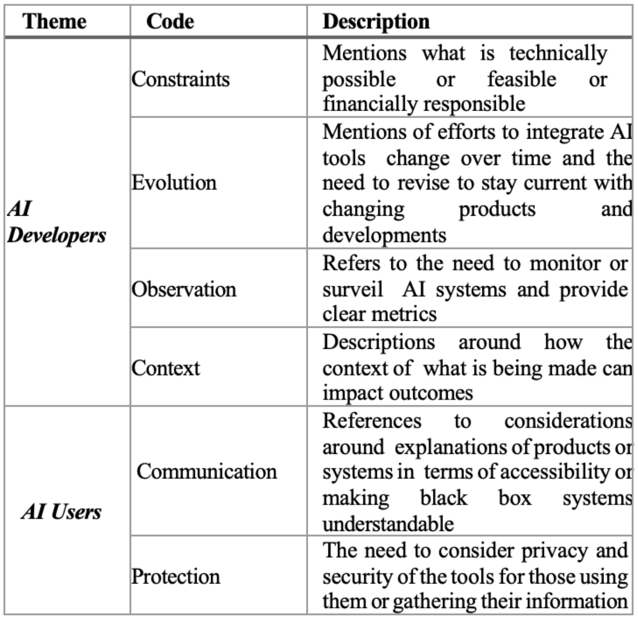
Figure 1 displays the countries assessed in this study and the respective documents included in the analysis. It also presents the year of publication for each.



Rank	Document Title	Year
USA	AI Ethics Framework for The Intelligence Community	2020
China	Governance Principles for the New Generation AI--Developing Responsible AI	2019
UK	Ethics, Transparency and Accountability Framework for Automated Decision-Making	2023
Israel	Responsible Innovation: Israel's Policy on AI Regulation and Ethics	2023
Canada	Responsible use of AI	2023
France	Transparency and accountability: the challenges of AI	2020
India	Responsible AI #AIFORALL	2021
Japan	Governance Guidelines for Implementation of AI Principles	2022
Germany	Ethics of AI	2020
Singapore	Model AI Governance Framework	2020

Figure 1: Top Ten Countries Leading in AI Ethics Development and Research according to Global AI Index Report 2023

According to the Global AI Index Report, these top ten countries have developed AI ethics frameworks and guidelines to govern the responsible use of artificial intelligence technologies. For example. The United States (US) AI Ethics Framework focuses on procuring, designing, and managing AI systems, drawing from an ethics guide for the US Intelligence Community. China's principles aim to promote healthy AI development and global collaboration. The United Kingdom (UK)'s principles aid government departments in ethical decision-making regarding automated systems. Israel's policy on AI regulation and ethics is geared towards advancements in the private sector and focuses on responsible AI innovation. Canada follows a shared approach to responsible AI with Digital Nations. France prioritizes investment, sovereignty, and ethics in AI regulation. India's guidelines assess AI risks, and legislation practices, and recommend responsible management principles. Japan's principles apply to society and advocate for goal adherence by AI companies. Germany aligns its AI ethics guidelines with the EU guidelines on ethics in AI. Singapore's AI Governance Framework incorporates international input for comprehensive regulation. Overall, these initiatives by all the nations mentioned above demonstrate a global effort to address ethical concerns and ensure AI technologies' safe and beneficial use.



Theme	Code	Description
AI Developers	Constraints	Mentions what is technically possible or feasible or financially responsible
	Evolution	Mentions of efforts to integrate AI tools change over time and the need to revise to stay current with changing products and developments
	Observation	Refers to the need to monitor or surveil AI systems and provide clear metrics
	Context	Descriptions around how the context of what is being made can impact outcomes
AI Users	Communication	References to considerations around explanations of products or systems in terms of accessibility or making black box systems understandable
	Protection	The need to consider privacy and security of the tools for those using them or gathering their information

Figure 2: Example of Coding System used in this study.

Thematic Analysis

Thematic analysis is a method for finding, evaluating, and summarizing patterns or themes in data (Braun & Clarke, 2006). This study focuses on the explainability and transparency components of the identified guidelines. This means that the paragraphs and sections in the documents that mention the terms explainability and transparency were the focal areas for analysis. These sections or paragraphs were placed into a dataset for analysis. The word count from the sections on transparency and explainability in the order of the country rank is 171, 93, 142, 199, 73, 197, 107, 86, 122, and 601, which helps to display details on the corpus size of the dataset. The thematic analysis was conducted using procedures previously described by (Braun & Clarke, 2012) to determine how these ten AI ethics guidelines discussed transparency and explainability. To enhance reliability, codes, and themes were generated separately by two independent raters based on patterns found in the guidelines. The raters then met to negotiate and finalize the codebook before coding.

Findings

This study aimed to determine how these ten AI ethics guidelines from these countries discussed transparency and explainability. The themes generated allowed for addressing this aim. There were two general themes for which these guidelines were classified. The complete list of codes and themes is shared in Figure 2.

a) AI Developers

We noted that the guidelines described reflected considerations for the "AI tool developers" theme, recognizing those who may be involved in their development and maintenance in multiple ways. It highlighted the various constraints developers face, often not known by non-AI experts. These included the feasibility and financial costs attached to implementing ethical AI. This was exemplified in an excerpt stating:

"...At the same time, the imposition of broad explainability requirements might be technically complex and financially onerous, potentially inhibiting innovation... " – (Responsible Innovation: Israel's Policy on AI Regulation and Ethics, 2023)

Additionally, these guidelines acknowledged that ethical AI practices are not stagnant. Therefore, the guidelines allow AI developers to continue evolving the tools and meet technological advances. For example, the following was mentioned in the guidelines:

"...We study the AI ethics discourse in other institutions, organizations, and companies to check and improve our guidelines to avoid a gap between theory and practice..." - (Ethics of AI, 2020)

Other guideline implications for AI Developers were observation or surveillance of AI systems to ensure protections like robustness, traceability, regular tuning, reproducibility, and controllability. As was noted:

"...The transparency, interpretability, reliability, and controllability of AI systems should be improved continuously to make the systems more traceable, trustworthy, and easier to audit and monitor..." - (Governance Principles for the New Generation AI-Developing Responsible AI, 2019)

Lastly, and importantly, the guidelines highlighted that AI tool developers should be aware of context as it could dramatically impact the outcomes of automated decisions. Another guideline noted:

"...Context is essential to the explainability of an automated decision... " - (Ethics, Transparency and Account-

ability Framework for Automated Decision-Making, 2023)

b) AI Users

The "AI Users" theme referred to factors for non-AI developers who interact with AI systems. The guidelines discussed communication and clarity as important for AI Users, with reference to how they may struggle to understand the black box of such tools and make it understandable. Examples from the guidelines are:

"... The explanation needs to be appropriate for your audience, expert or non-expert and should be scrutinized and iterated by a multidisciplinary and diverse team (including end users) to avoid bias and group speak. . . " - (Ethics, Transparency and Accountability Framework for Automated Decision-Making, 2023)

"...In this context, "explainability" refers to the ability to explain how a particular AI system operates or to provide the reasons for a specific AI-based decision or recommendation, in a manner that is readily understandable..." - (Responsible Innovation: Israel's Policy on AI Regulation and Ethics, 2023)

In addition, protection for end users was prioritized in terms of the privacy and security of their information. Examples from the guidelines are:

"...the Commission's White Paper calls for an "ecosystem of trust" to ensure the protection of fundamental rights, security, and regulatory stability. . . " - (Transparency and accountability: the challenges of AI, 2020)

". . . Companies are encouraged to test, evaluate and review their strategies for effectiveness. . . " = (Model AI Governance Framework, 2020)

These findings, though based only on transparency and explainability, suggest more profound ways to view their importance from an international policy standpoint.

Discussion

The results highlight deeper insights from just the transparency and explainability portions of global AI ethics guidelines. It is very easy to rule out transparency and explainability as entirely technical concepts, but taking an international collective perspective enhanced the nuances of transparency and explainability, encapsulating the many areas in which it manifests in policy documents. The findings testify to the importance of taking a global and collective approach to AI ethics. As described by the two themes "AI Developers" and "AI Users," much of what is presented in these guidelines coincides with existing literature. For instance, we observe that even though XAI is desired, it can be costly or even not feasible to implement in some cases, as Mc Dermid et al. (2021) noted. In addition, transparency and explainability are, in fact, sociotechnical concepts having layers in both the technical and social domains Ehsan et al. (2021).

From this international perspective, countries developing policies may apply these insights to ensure effective communication with all AI stakeholders. Explainability was more commonly discussed in the policies than transparency. This can be attributed to the fact that transparency is an umbrella term that comprises explainability. Additionally, the results highlight that explainability is a critical area in AI ethics for AI developers and AI users as it encompasses rationale for decisions made by AI. For AI developers, it implies a need to continue improve this area through practice and research, not only from a technical standpoint but from a multidisciplinary one. For AI users, explainability helps in developing trust in AI systems. There are also implications for academia in preparing students to take roles as AI developers or AI application users. It is critical for AI curriculum developers to ensure that both technical and social dimensions of transparency and explainability are reflected in the curriculum with readiness to ensure students are fully prepared to handle the global and industrial challenges AI may pose. It also highlights the need for more interdisciplinary collaboration within different departments at the university level and across public and private sectors. Therefore, collectively analyzing ethical AI guidelines can be insightful as it helps in understanding stakeholders.

Limitations

There are several limitations we want to acknowledge. First, the policy documents studied only represent a subset of all those that may exist in the entire global landscape of AI ethics. Other countries not examined may take different approaches. Furthermore, some of the documents explored were geared towards the country's national AI strategy and were not specifically solely for responsible AI or AI ethics. Additionally, the researchers used the documents available in English, which may limit what information can be gleaned. As a result of disparities in text lengths among the documents, the items of interest (transparency and explainability) were unevenly represented, leading to differing amounts of data available for analysis from each document.

Conclusion

When analyzing AI ethics guidelines, it can be beneficial to take a global approach as it widens the understanding of viewing AI ethics. Combining guidelines from leading AI ethics nations enabled a thorough review of transparency and explainability, giving insights into important details of the technicality of explainable AI that can be easily overlooked by nontechnical AI stakeholders. The thematic analysis proved helpful in developing themes to understand better the phenomena studied for both AI Creators and users. A comprehensive take on transparency and explainability reveals these often-called technical dimensions to possess more sociotechnical layers involvinh AI and non-AI developers. Going forward, we hope findings will inspire AI stakeholders, policymakers worldwide, as well as academic curriculum developers, to convey the sociotechnical layers of transparency and explainability of AI.

Reference

1. Ad Hoc Expert Group. (2020). Outcome document: First draft of the recommendation on the ethics of artificial intelligence.

2. Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022, March). Transparency and explainability of AI systems: ethical guidelines in practice. In International Working Conference on Requirements Engineering: Foundation for Software Quality (pp. 3-18). Cham: Springer International Publishing.

3. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

4. Braun, V., & Clarke, V. (2012). Thematic analysis. American Psychological Association.

5. Cesareo, S., & White, J. (2023). The Global AI Index. Tortoise.https://www.tortoisemedia.com/intelligence/global-ai/#rankings

6. Deshpande, A., & Sharp, H. (2022, July). Responsible AI Systems: Who are the Stakeholders?. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 227-236)

7. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021, May). Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-19).

8. Ethics Council: Artificial intelligence must not diminish human flourishing. (2017, January 21). Www.ethikrat.org. https://www.ethikrat.org/en/press-releases/press-releases/2023/ethics-council-artificial-intelligence-must-not- diminish- human-flourishing/?cookieLevel=not-set#: :text=On%2020%20March%202023%2C%20the

9. Ethics, Transparency and Accountability Framework for Automated Decision-Making. (2021). GOV.UK. https://www.gov.uk/government/publications/ethics-trans- parency-and-accountability-framework-for-automated-de- cision-making/ethics-transparency-and-accountability- framework-for-automated-decision-making étrangères, M. de l'Europe et des A. (n.d.).

10. Transparency and accountability: the challenges of artificial intelligence. France Diplomacy - Ministry for Europe and Foreign Affairs. https://www.diplomatie.gouv.fr/en/french-foreign- policy/digital-diplomacy/transparency-and-accountability- the-challenges-of-artificial-intelligence/

11. Expert Group. (2022). Governance Guidelines for Implementation of AI Principles. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

12. Guo Kai. (2019, June 17). Governance Principles for the New Generation Artificial Intelligence–Developing Re- sponsible Artificial Intelligence. Www.chinadaily.com.cn. https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html

13. Jonas Bedford-Strohm, U. K. (2022). Ethics of Artificial Intelligence: Our AI Ethics Guidelines. Www.br.de. https://www.br.de/extra/ai-automation-lab-english/ai-ethics100.html

14. Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2).

15. Lawton, G. (2023, June 2). AI transparency: What is it and why do we need it? Tech Target - CIO. https://www.techtarget.com/searchcio/tip/AI-transparency-What-is-it-and-why-do-we-need-it

16. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45

17. McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. Philosophical Transactions of the Royal Society A, 379(2207), 20200363.

18. Ministry of Innovation, Science and Technology. (2023). Responsible Innovation: Israel's Policy on Artificial Intelligence Regulation and Ethics. https://www.gov.il/BlobFolder/news/most-news20231218/en/Is- raels

19. OECD.AI (2021), powered by EC/OECD (2021), database of national AI policies, https://oecd.ai.

20. Office of the Director of National Intelligence, & Admin. (n.d.). Artificial Intelligence Ethics Framework for the Intelligence Community. INTEL. https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the- intelligence-community#Intro

21. Roy, A. (2021). Approach Document for India Part 1 -Principles for Responsible AI FEBRUARY 2021 RESPONSIBLE AI #AIFORALL. https://www.niti.gov.in/sites/de- fault/files/2021-02/Responsible-AI-22022021.pdf

22. S Iswaran. (2020). ARTIFICIAL INTELLIGENCE GOV- ERNANCE FRAMEWORK MODEL SECOND EDITION.https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/re- source-for-organisation/ai/sgmodelaigovframework2.pdf

23. Schmitt, L. (2022). Mapping global AI governance: a nascent regime in a fragmented landscape. AI and Ethics, 2(2), 303-314.

24. Secretariat, T. B. of C. (2023, December 13). Government of Canada. Canada.ca. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1

25. Woudstra, F. (2020). Ethical Guidelines for Transparent Development and Implementation of AI-an Overview. Filoso- fie in Actie blog.