TRANSPARENCY AND EXPLAINABILITY THEMATIC AREA NARRATIVE IN ENGLISH ARABIC FRENCH PORTUGUESE AND SPANISH

Rachel Adams , Kelly Stone

Rachel Adams , Kelly Stone

IDRC GRANT / SUBVENTION DU CRDI : - GLOBAL INDEX ON RESPONSIBLE ARTIFICIAL INTELLIGENCE

Global Index on Responsible AI

Dimension: Responsible AI Governance

Sub-dimension: Rule of Law

Thematic area: Transparency and Explainability

Definitions

Transparency is defined as 'the quality of [a process, practice or decision] being done in an open way without secrets so people can trust that [actions] are fair and honest'. Transparency has been identified by the UN Human Rights Council (UNHRC) as one of the five key principles of good governance, along with responsibility, accountability, participation and responsiveness to public concerns.

In the context of AI, transparency concerns the extent to which the "inner workings" of an AI system are open and accessible, which invariably requires the provision of "easy-to-understand" explanations of the algorithmic models, the data that drives them, and the rationale for their use. It also refers to transparency in the use of an AI system, whether by a public or private actor, and particularly toward people who are likely to be impacted by the use of such a system. To this end, transparency aims to ensure that people are able to develop a basic understanding of AI systems, including the reasons for their development, training, deployment, and impact.

Accordingly, the principle of transparency implicates the concept of explainability, which has arisen within the AI ethics discourse and which calls for the provision of information that can be understood by a non-technical person – particularly, the process driving outcomes of AI systems. Explainability can be defined as 'the quality of enabling people to understand how a particular system works and/or how a particular outcome was achieved by providing information that is sensible and easy-to-understand'. The principle of explainability is also closely linked to the concept of interpretability, which refers to 'the capacity to provide an explanation of the factors and logic that led to an outcome in terms that are understandable to a human.'

Transparency and explainability, as principles of responsible AI governance, therefore require information about processes, practices and decisions to be open, accessible, and easily understood by a broad range of stakeholders, including government entities, private sector companies, civil society organisations and academic institutions.

Justification

AI systems and tools are continuously evolving and increasingly permeating today's digital landscape and society at large. Reliance on AI technologies to perform various functions both with and without people's knowledge, such as making predictions, recommendations, and decisions, has reinforced the need for increased levels of transparency around use of these systems and explanations for how they work and the impact they make. Understanding the reasoning behind AI-generated decisions, the timing and purpose of deploying AI systems, as well as the processes and training data that shape them, becomes crucial due to the potential risks associated with these emerging technologies.

The UNESCO Recommendation on the Ethics of AI (Recommendation) describes transparency as a 'necessary precondition for the respect, protection and promotion of human rights'. The Recommendation acknowledges that a lack of transparency can 'undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems', which can impact on the right to a fair trial, access to remedy and redress, as well as the areas and industries in which AI systems can be used. The Recommendation also notes that greater levels of transparency can 'enable people to understand how each stage of an AI system is put in place' which may identify factors that affect a specific prediction or decision and highlight whether or not procedural safeguards are in place, such as safety or fairness measures.

Further, the Organisation for Economic Development and Cooperation (OECD) notes that the principle of transparency requires disclosure around the use of AI systems, specifically when AI tools are making predictions or recommendations, or when people are directly engaging with AI agents. Transparency requires people to 'understand how an AI system is developed, trained, operates and is deployed' by providing access to relevant information that is easy to understand so people can make informed decisions. Further, the OECD also notes that a critical element of transparency is providing mechanisms for facilitating open dialogues amongst stakeholders and establishing dedicated entities/institutions to promote general awareness of AI, to respond to public concerns over its use, and to increase public confidence, trust, and acceptance of its responsible application.

Identification

This thematic area measures steps countries have taken to promote and ensure sufficient levels of transparency and explainability in the development, use and deployment of AI technologies. In particular, evidence must account for (1) frameworks concerning standards of transparency and explainability in AI use, development, and deployment, (2) government actions to promote the principles of transparency and explainability in the use of AI systems, and (3) non-state actors

working to advance, enable or enforce transparency and explainability throughout all phases of the AI lifecycle.

Frameworks may take the form of laws, regulations, policies (including by sector and/or department) and/or guidelines. Government actions may include draft laws or policies, the establishment of oversight bodies to monitor compliance, or the implementation of policies that seek to make AI systems more transparent and explainable. Non-state actors (NSAs) may include non-governmental organisations (NGOs), but also multinational corporations, private military organisations, media outlets, organised ethnic groups, academic institutions, lobby groups, labour unions or social movements that are working to advance the transparency and explainability of AI systems.

Examples

Frameworks

Australia has established a set of AI Ethics Principles to guide the responsible and ethical development and use of artificial intelligence. The principles provide a framework for individuals and organizations to ensure that AI technologies are used in a manner that aligns with Australia's values and societal expectations. One of the core principles include Transparency and Explainability, which states 'there should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.' More specifically, the principles assert that responsible disclosures should be timely and 'provide reasonable justifications for AI systems outcomes', which necessarily includes information that helps people understand the outcomes, including the factors considered in decision-making processes. These principles are meant to guide the responsible and ethical use of AI technology across various sectors in Australia, promoting trust, fairness, and societal benefits

Government Actions

In April 2021, the European Commission released the Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence, (AI Act). The AI Act classifies different types of technologies according to three levels of risk (unacceptable, high-risk, and low or minimal) and then prescribes certain obligations on AI providers to mitigate potential harms. Title IV places specific transparency obligations on systems classified as high-risk, including those that: (1) interact with humans; (2) be used to detect emotions or determine association with (social) categories based on biometric data; or (3) generate or manipulate content ('deep fakes'). For example, when individuals interact with an AI system or when their emotions or characteristics are recognised using automated means, Title IV requires users of AI to *inform* people they are being subjected to that technology and *explain* how it works. In addition, Title IV requires users of AI systems to disclose when content is being generated using automated means, unless it falls

under one of the prescribed exemptions in law, to ensure persons have the opportunity to make informed choices about their level of engagement.

While the AI Act has yet to come into effect, significant efforts are underway to ensure that both the European Council (representing the 27 EU Member States) and the European Parliament agree on a common version of the text. On 6 December 2022, the Council adopted its common position ('general approach') on the AI Act, which it will use to engage in negotiations with the European Parliament before coming to an agreement. According to recent reports, the AI Act is expected to be signed into law by the end of 2023, after which time all providers of AI systems will be expected to comply within 24-36 months.

Non-Government Actors

Credo AI, based in the United States, is one of numerous companies specialising in governance, risk and compliance (GRC) software platforms empowering businesses to responsibly manage AI assets, offers an innovative 'AI Governance Platform' to help businesses "analyse, audit and manage the risks" posed by machine learning and AI systems. In November 2022, the company released the feature of new assessment and reporting capabilities to further facilitate enterprises' compliance with regulatory obligations as well as consumer demands for "governance artifacts, reports and disclosures" around their development and use of AI with specific regard to explainability alongside fairness, robustness, security and privacy concerns.

Research Guidance

With regard to relevant frameworks, start by identifying any binding law or regulation that aims to govern AI. Note the different terms often synonymous with transparency regarding AI such as 'notice', 'disclosure', and 'black box'. Beyond generic references to transparency in AI use and development, look for evidence of specific regulatory requirements that call for notice requirements or specific policy recommendations in relation to transparency. Transparency requirements may be proportional to perceived risk levels. AI systems may also be referred to as 'algorithmic systems', and it may be useful to include 'machine learning' in searches.

In relation to government actions, start by searching the responsible ministry website (e.g. Ministry of Economics, Ministry of Technology, Ministry of Justice, etc.) and identify events, monitoring mechanisms or information related to transparency and explainability. Check if there might be guidelines specifying the application of general laws (consumer protection etc.) to the use-case of AI that impose specific notice requirements or dissemination specifications to ensure adequate explainability.

To identify non-state actors active in AI transparency and explainability, start by looking at national consumer advice centres and identify any work they have done in relation to notice/disclosure or dissemination requirements. Moreover, national civil organisations working on digital policy, data protection and/or human rights will be

worth investigating, but make sure these activities are specific to the promotion of transparent and explainable AI systems.

Some Useful Sources

- Existing literature on the (complex) issue of transparency around AI/algorithmic systems, including new developments/reforms in the country and recent academic research (e.g. reports, policy briefs, news/articles, white papers, academic papers)
- Websites of civil society/other non-governmental organisations promoting the issue of or addressing transparency and/or explainability in AI
- OECD.AI (live repository of global AI policy initiatives)

Search

- Parliamentary or government records for recent mentions of 'transparency and AI', 'disclosure and AI', 'explainable AI' etc.
- General google searches for 'transparency and AI and [country]', 'explainability and AI and [country]', etc. (can also include terms 'policy', 'framework', 'law', etc.)
- Academic search engines (Google Scholar, arXiv, ResearchGate, etc.) for papers on 'transparency and AI and [country]', 'explainability and AI and [country]', etc.

Consult

- Civil society organisations in the country dealing with AI transparency and explainability such as disclosure, public participation, etc.
- Academics/researchers specialising on transparency and explainable AI
- Computer scientists developing or AI experts regarding frameworks/solutions supporting AI transparency/explainability
- Businesses or corporations providing tools/resources in support of transparent and explainable AI