The Hague International Model United Nations

Forum: General Assembly 6 (Legal)

Issue: Measures for the transparency, accountability, and oversight of Artificial Intelligence Systems

Student Officer: Aarav K Reddy

Position: Deputy Chair

Introduction

Artificial Intelligence systems, henceforth referred to as 'AI', are increasingly integral segments of our daily lives. They successfully analyze data, automate complex processes, and enhance decision-making across countless pivotal industries, such as healthcare, finance, and education, that cannot afford to be negatively impacted by any of AI's potential slip-ups (including risks due to overfitting on training data, a lack of data security, or lack of data for a particular situation causing AI to give a wrong result with high confidence). Even considering that AI was only recently introduced to modern-day living, the international regulation of AI systems is severely behind the rapid advances of such technologies. Given the opacity of the internal functioning (i.e. what models, hyperparameters, and training data are used) of AI systems, alongside their exponential development and usage, the lack of regulation is a global issue (for example, 83% of companies claim AI as a top priority, and the use of AI in the workforce jumped 270% over four years as of 2019). Hence, it is supremely necessary to work on setting policies in place that address factors such as transparency, accountability, and oversight measures, even if it is hard to find the line between stumping innovation and protecting humanity. .

The lack of AI's transparency raises ethical and societal questions with far-reaching implications, especially as these systems seem increasingly poised to influence critical areas such as criminal justice, lending, and public policy. Without any clear accountability mechanisms, there is a risk of harm due to biased algorithms, flawed decision-making, and potential misuse of AI technology.

Moreover, AI systems' rapid deployment has outpaced the establishment of robust oversight frameworks, with only 29% of businesses correctly implementing safe AI practices, despite 78% believing they do.

AI's integration into our governance and decision-making is sure to increase. It can be seen as an incredible tool, however, it must be noted that the potential for misuse and systemic harm conducted with it increases the more industries and daily tasks it is used in. Addressing these challenges through clear oversight structures, alongside accountability and transparency measures on an international level, is vital to ensuring that AI serves our global society responsibly and equitably.

Definition of Key Terms

Accountability

The legal frameworks, regulations, and processes that hold the general population, alongside private actors, and public officials legally responsible for any actions they may perform, and that impose legal consequences if the law is found to be violated by any party.

Artificial Intelligence

Artificial Intelligence (AI) refers to machines or software demonstrating intelligent behavior, akin to the capacity for reasoning in humans. It refers to the field of study in computer science that develops and studies such intelligent machines or the intelligent machines themselves.

Data Privacy

The principle of data privacy is that an individual or collective should have control over their personal data, including the ability to decide how other organizations collect, store and use their data. This includes safeguarding sensitive data from unauthorized access and ensuring compliance with privacy regulations.

Ethical Implication

The potential ethical considerations and consequences that arise from research or professional aspects of that activity, such as issues related to objectivity and bias, alongside other ethical practices like copyright and plagiarism.

Oversight

The general process of review, monitoring, evaluation, supervision, reporting and audit programmes, activities, policy implementation, and results of an organization or technology.

Surveillance

The non-consensual interception of communications or private data, electronic or otherwise, the collection, storage and processing of that data for personal benefit, and transfer of that data to third parties.

Transparency

The sharing of data that is of clear public interest, including operations management, governing rules, activities, expenditures, and product design, to the public or to a concerned regulatory authority.

Background Information

Transparency in AI Decision-Making:

Despite the importance of AI in key decisions, from healthcare and criminal justice, to financial lending, there is a stark lack of trust in the public about its efficacy, and how it is used by corporations and decision-makers. This lack of transparency has its roots in three key regions, leading to unfair bias, unjust discrimination, and a general weakening of public trust in AI technologies that cannot be resolved due to a lack of both public and legislative awareness.

There are three key issues stemming from this opacity. Most evidently, the unregulated discrimination and systemic bias showcased by certain AI algorithms, evidenced in how hiring algorithms exhibit gender and racial biases or deny opportunities to marginalized groups. For example, many computer-aided diagnosis (CAD) systems are shown to return lower accuracy results for black patients than white patients. This occurs as a result of data being fed to the algorithms that is, in itself, extracted from biassed sources– and which can contribute to the long-term entrenchment of these biassed practices. Another example includes when researchers at Carnegie Mellon University in Pittsburgh revealed that Google's online advertising system displayed high-paying positions to men more often than to women. These factors result in a second issue, a loss of trust among the public: in regulators, corporations, and AI systems themselves, overshadowing their perceived benefits and undermining their credibility. Finally, both these issues contribute to the larger problem of the erosion of legal and moral responsibility in companies/developers, who may feel emboldened due to the difficulty of holding them accountable for AI-related mistakes, ultimately sowing a culture of distrust that leaves reduces the avenues of justice available for those negatively affected by extralegal uses of AI.

The above three issues originate from three characteristic origins. Firstly, the complexity of AI algorithms, namely the deep learning and machine learning models, ensures reduced transparency among business executives, governors, and the public. This is exacerbated by AI systems solely being able to learn from historical data, which can contain ingrained biases. If transparency mechanisms are not built into the AI's decision-making processes, these biases go unrecognized and perpetuate unfairness. Furthermore, the absence of uniform standards for AI accountability has led to diverse and inconsistent approaches to transparency. In many jurisdictions, companies operate without clear oversight, allowing them to keep AI processes confidential without scrutiny. Lastly, many companies prioritize technological innovation and profitability over ethical responsibility. As a result, companies are more focused on the development of AI systems rather than on ensuring these systems operate transparently and ethically.

Though the issue of AI transparency has recently entered the public eye, there have been major developments both now and in the past. For example, Explainable AI (XAI) was invented to relieve concerns about machine learning models' opacity, especially that of neural networks. Researchers began developing techniques to make AI systems interpretable, such as feature attribution methods

(e.g., Shapley Additive Explanations and Local Interpretable Model-agnostic Explanations, both in 2016) and visualization tools. These approaches carry out decisions while also explaining how AI systems reach them, allowing stakeholders to evaluate the fairness and accuracy of these processes. Furthermore, some corporations themselves have promoted AI transparency. Namely, Google's AI principles, announced in 2018, included a commitment to transparency by providing explanations for AI decisions and making clear when AI is in use, and IBM's AI Fairness 360 toolkit, which works to help developers identify and mitigate bias in their models, promoting interpretability and fairness.

Regulatory Ethical Frameworks for AI Governance

Establishing regulatory ethical frameworks for AI governance is necessary to ensure the transparency, accountability, and oversight of AI systems remains entrenched in law. These frameworks work to form clear guidelines that propagate fairness, address bias, and protect human rights. This ensures an alignment between ethics/societal norms and the AI's functioning. Yet, the development of these frameworks has encountered challenges ranging from technological advancements outpacing legal systems to varying international approaches to AI oversight.

It is necessary to create these frameworks specifically as the AI market is expected to grow past 826 billion USD by 2030. Proper governance on the international level would ensure sustainable growth and ethical deployment of AI while also helping harmonise regulations across borders, to prevent inequity or an 'AI-development race', reducing conflicts arising from differing national approaches and standards. They would additionally protect marginalized groups from unfair treatment by automated systems, critical in the applications already mentioned (eg: sensitive applications like healthcare, finance, and criminal justice).

Steps toward formal AI regulation were first properly initiated in the 1990s, with prominent local preliminary regulations like the European Union's Data Protection Directive, which began in 1995 to address issues indirectly related to AI, like data privacy and the ethical use of personal information. At the same time, research institutions' and corporations' internal ethics boards to evaluate AI projects grew in prevalence but these lacked formal enforcement mechanisms. In the 2010s, high-profile incidents like biased hiring algorithms and discriminatory AI systems led to calls for transparency (a 2015 study revealed racial bias in predictive policing algorithms). Thus, certain key AI regulations were introduced. These included the General Data Protection Regulation (GDPR) in 2018, which addressed data privacy and transparency to a bigger degree than before, requiring organizations to disclose how AI systems use personal data. The regulation further introduced the concept of a "right to explanation" for decisions made by automated systems, helping with AI transparency. Furthermore, the Asilomar AI Principles (2017) outlined 23 ideal guidelines for AI development and ethics, which were endorsed by leading AI researchers and tech companies. Additionally, the IEEE Global Initiative on Ethics of

Autonomous and Intelligent Systems (2016) provided standards for ethical AI design. However, neither has been applied on the national level for any country yet nor on the international level.

On the other hand, international initiatives like the OECD principles and the UN's ethics of AI do provide a baseline set of rules for ethical AI use (though these are less broad than local programmes), allowing organizations to adopt ethical practices without rigid enforcement. They also encourage industries to consider fairness, transparency, and accountability– all of which are like cornerstones of positive long-term AI policies. However, due to their non-binding nature which results in an absence of enforcement mechanisms, it is difficult to rely on parties to comply with regulations. Additionally, when applied, it generally has an uneven application across industries and countries. This results in certain countries that refuse to restrict themselves from having an advantage over others, resulting in countries receiving a form of reward for non-compliance, further encouraging that behaviour.

Currently, two key issues prevent most international regulatory advances. These include global fragmentation– specifically how national priorities and regulatory approaches vary significantly across regions, leading to inconsistencies when it comes to multinational regulation. Secondly, the high pace of innovation causes regulations to struggle to keep up with rapidly advancing AI capabilities, such as generative AI models like GPT or DALL-E.

Major Countries and Organizations Involved

United States of America

The US has been instrumental in setting benchmarks for AI ethics and regulatory frameworks through efforts like the Blueprint for an AI Bill of Rights released by the White House Office of Science and Technology Policy (OSTP), establishing key transparency, accountability, and anti-discrimination measures. Additionally, it has balanced its own regulatory initiatives with global standards (including the EU's AI Act and the OECD's AI Principles). Furthermore, the US also has a policy of advocating for open, transparent, and ethical AI systems, balancing that with a progress-focussed agenda that seeks consistent AI innovation to ensure America's leading role in this sector. It also works on nationally standardising technical frameworks to ensure the trustworthiness and security of AI systems, alongside collaborating with industry leaders, universities and think tanks, including OpenAI, Google, and Microsoft, to ensure that commercial and research AI systems adhere to ethical and regulatory standards. The US additionally seeks to prioritize governmental focus on the ethical and socially beneficial applications of AI, allocating federal funding for such projects.

Ultimately, the US approach to AI governance reflects a general commitment to prioritising citizens' rights, encouraging competition and maintaining the country's global influence and dominance in the AI sector. However, the USA's participation in international or regional agreements for AI is

relatively minimal, limiting the influence and curtailment that intergovernmental organisations can have on the country.

China

China's lack of transparency in AI development is one of the biggest concerns lobbied toward it. Moreover, the use of Chinese AI in social governance (namely via the Social Credit System) acts as a potential misuse of data for political or social control, flouting principles stated through draft regulations like China's Generative AI Rules, which require transparency in algorithmic development, data usage, and content generation. Furthermore, Chinese companies like Huawei and Hikvision export AI-based surveillance technologies to countries with weak regulatory frameworks, potentially enabling human rights abuses and violating China's own "Beijing AI Principles" that emphasize human well-being, fairness, and sustainability in the use of AI.

On the other hand, China is actively involved in multilateral forums such as UNESCO and the Global Partnership on Artificial Intelligence (GPAI), alongside working with developing countries on AI use. With initiatives like the Belt and Road Initiative (BRI) in those countries, China promotes the use of AI for social and economic development. Consequently, this engagement in such platforms implies a desire to contribute to the large-scale global standards for ethical AI use. On the other hand, China's strategic focus on dominating the AI sector has led to stark competition with the global West, namely the US and the EU, which can hinder international cooperation on ethical frameworks and oversight mechanisms in the long run, and also lead to conflict.

UN

The UN's position as a multilateral organization enables the creation of universal principles that can be adopted worldwide, reducing regulatory fragmentation which is a key issue. However, this hasn't occurred yet due to "universal ethical principles" perhaps conflicting with national laws, cultural values, or existing regulatory priorities, leading to uneven adoption. Furthermore, it leads to issues regarding the actual enforcement of these regulations, which is difficult to do, particularly for states with more decentralised institutions, corruption, or a lack of interest in following through on regulations. This has led to most of the UN's steps taken with this agenda being voluntary, as anything beyond that is considered infeasible.

Furthermore, AI is a rapidly developing industry. The multilateral nature of the UN can fall victim to that, due to its laws' prolonged deliberations and delayed implementation, which are likely to fail to keep pace with rapidly advancing AI technologies. Furthermore, wealthier nations that are influential over the UN and large tech corporations that are influential over those countries may dominate the agenda-setting process, potentially sidelining the interests of smaller or less-developed nations and reducing the equity proffered by potential legislation. However, the UN is currently the only international

medium through which every developing nation can be sufficiently aided in growing in the AI era, its far reach and relatively higher legitimacy compared to alternatives necessitates the passing of policies through it. However, if this legislation leads to overregulation, we may have the opposite problem that we do today. Too-restrictive ethical and regulatory frameworks could easily hinder AI innovation and discourage investment in new technologies, stagnating the sector before it is allowed to prosper, which would lead to the effect of reducing the use of AI, unintended by all countries parties to the UN.

Timeline of Events

Date	Name	Description of Event
December, 13th, 1995	Entry into force of the European Union's Data Protection Directive	One of the first pieces of regulation affecting AI, addressing issues like data privacy and the ethical use of personal information.
April, 2016	Launching of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	To incorporate ethical aspects of human well-being in AI development.
August, 10th, 2016	The DARPA explainable AI initiative is created	This initiative works to make AI less of a "black box" and increase the transparency of AI
June, 7th, 2017	Establishment of the 'AI for Good' summit by the International Telecommunication Union (ITU)	Established by the ITU, the UN agency for digital technologies, the platform's mission is to use AI to help achieve the UN SDGs.
May, 25th, 2018	The EU general data protection regulation enters into application.	The EU General Data Protection Regulation (GDPR) is the strongest privacy and security law in the world. This regulation updated and modernised the principles of the 1995 data protection directive
January, 5th, 2021	Public release of DALL-E	DALL-E is the most widely used consumer image-generator AI.
September, 20th, 2022	The Principles for the Ethical Use of Artificial Intelligence in the United Nations System are released	This is a landmark document by the UN suggesting guidelines for AI use internationally
November, 30th, 2022	ChatGPT was first released to the public	ChatGPT is the most widely used consumer text-generator AI.

May, 3rd, 2024	The Updated OECD AI Principles are released	They help countries respond to risks emerging from the latest technological developments such as general-purpose and generative AI systems
August, 1st, 2024	EU AI Act enters into force	The Act aims to foster responsible artificial intelligence development and deployment in the EU and acts as a guideline for the rest of the world.

Previous Attempts to Solve the Issue

The European Union AI Act is an important attempt to look at. Its advantages include its risk-based framework, which ensures resources are focused on systems with the greatest potential for harm, focussing overseers' efforts and establishing clear governance roles for oversight. However, there is a potential that its classification is too rigid, not accounting for evolving technologies. Furthermore, the difficulty it makes for companies to comply with its complex, high-cost laws can stifle innovation and lead to uneven implementation across member parties.

The Organisation for Economic Co-operation and Development's (OECD's) AI Principles have the positives of having a high legitimacy due to the OECD's soft power, namely its good reputation and the organisation's broad membership. However, its non-binding nature means it has limited enforcement power, which reduces the legislation's legitimacy, as does its assumption that all countries have the resources to fulfil its conditions– making it very difficult to ensure consistent adoption across countries with varying resources. Further negatively affecting localisation efforts is the vagueness of national-level implementation details.

As for initiatives within the industry, Google's AI Model Cards help increase transparency by providing clear documentation of models' purposes and limitations while also encouraging industry-wide adoption of similar documentation practices. This helps both developers and users make informed decisions about what model to use. However, the high technical complexity of the cards is geared solely to technical users, making them less accessible to general stakeholders. Furthermore, it has minimal policies addressing accountability or oversight measures beyond disclosure of practices. So, although the companies may be transparent about negative practices, there is nothing stopping them from conducting those practices.

Possible Solutions

Currently, most issues hinge on finding a balance between the corporate desire to expand the use of AI in an unregulated manner, and citizens' desires to have further regulation in the industry to make it more transparent and improve data security, allowing AI to ethically grow.

A possible solution could thus include greater investment in explainable AI, with transparency audits identifying deliberately hidden gaps or key missing areas of information in AI systems. This could be paired with model documentation standards to explain how models and datasets were built, their intended use, and known limitations alongside either open sharing of datasets with the public or private sharing with the government to verify that consumer private data is not being stolen.

Furthermore, clearly defining lines of accountability for AI, including the roles of developers, deployers, and end-users, would help create greater compliance as companies know what rules and guidelines to adhere to. This could be achieved via a global adoption of policies like the EU's AI act, or merely nation-specific similar frameworks.

Moreover, the traceability of AI systems' decision-making process is essential to regain public trust in these algorithms, especially in case of key errors or damage caused by the system. This would further be strengthened through the establishment of legal frameworks that clarify liability for harm caused by AI, particularly in high-stakes areas like healthcare or autonomous vehicles along with some level of human oversight in these cases.

Lastly, for any of these procedures to have their intended beneficial effect, it is essential for there to be a legal enforcement system in place, either on the national or international level.

Bibliography

"AI for Good Global Summit 2017." *ITU*, www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx. Accessed 14 Dec. 2024.

"AI Principles." *OECD*, www.oecd.org/en/topics/sub-issues/ai-principles.html. Accessed 14 Dec. 2024.

Bianchi, Andrea. *On Power and Illusion: The Concept of Transparency in International Law*. assets.cambridge.org/97811070/21389/excerpt/9781107021389_excerpt.pdf.

Broad Agency Announcement. 2016, www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

"China Releases New Draft Regulations for Generative AI." *China Briefing News*, 30 May 2024, www.china-briefing.com/news/china-releases-new-draft-regulations-on-generative-ai/.

Cohen, Idit. "Explainable AI (XAI) with SHAP - Regression Problem." *Medium*, 23 Oct. 2021, towardsdatascience.com/explainable-ai-xai-with-shap-regression-problem-b2d63fdca670 .

Directorate-General for Communication. “AI Act Enters into Force - European Commission.” *European Commission*, 1 Aug. 2024, commission.europa.eu/news/ai-act-enters-force-2024-08-01_en.

“EUR-Lex - 31995L0046 - EN - EUR-Lex.” *Eur-Lex.europa.eu*, eur-lex.europa.eu/eli/dir/1995/46/oj.

European Council. “The General Data Protection Regulation.” *Www.consilium.europa.eu*, 13 June 2024, www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/.

EXECUTIVE SUMMARY INTERNATIONAL STANDARDS on TRANSPARENCY and ACCOUNTABILITY. 2014, www.law-democracy.org/live/wp-content/uploads/2014/04/Transparency-and-Accountability.final_.Mar14.pdf.

Future of Life Institute. “AI Principles.” *Future of Life Institute*, 11 Aug. 2017, futureoflife.org/open-letter/ai-principles/.

Gardhouse, Kathrin. “OECD AI Principles 2024: Addressing Generative AI New Risks.” *Private AI*, 12 June 2024, www.private-ai.com/en/2024/06/12/oecd-ai-principles-2024/. Accessed 14 Dec. 2024.

“Gartner: Enterprise Use of AI Grew 270% over the Past 4 Years.” *VentureBeat*, 21 Jan. 2019, venturebeat.com/ai/gartner-enterprise-ai-implementation-grew-270-over-the-past-four-years/.

GDPR. “General Data Protection Regulation (GDPR).” *General Data Protection Regulation (GDPR)*, 2024, gdpr-info.eu/.

Gilkes, Sarah, et al. “Navigating AI Regulation: Essential Insights for Australian Businesses.” *Lexology*, Hamilton Locke, 5 Dec. 2024, www.lexology.com/library/detail.aspx?g=bc725a5a-8762-4acf-9176-40ec7a33f81f. Accessed 11 Dec. 2024.

Google. “Google AI Principles.” *Google AI*, 2023, ai.google/responsibility/principles/.

“Google Model Cards.” *Withgoogle.com*, 2024, modelcards.withgoogle.com/. Accessed 14 Dec. 2024.

IBM. “What Is Explainable AI? | IBM.” *IBM*, 2024, www.ibm.com/topics/explainable-ai.

IBM Data and AI Team. “AI Bias Examples | IBM.” *Ibm.com*, IBM, 16 Oct. 2023, www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

“Introduction to Explainable AI(XAI) Using LIME.” *GeeksforGeeks*, 20 Jan. 2021, www.geeksforgeeks.org/introduction-to-explainable-aixai-using-lime/.

Milanovic, Marko. *Human Rights Treaties and Foreign Surveillance: Privacy in the Digital Age*. no. 1, 2015, www.ilsa.org/Jessup/Jessup16/Batch%202/MilanovicPrivacy.pdf.

National University. “131 AI Statistics and Trends (2024).” *National University*, 1 Mar. 2024, www.nu.edu/blog/ai-statistics-trends/.

“Oversight | United Nations Development Programme.” *Undp.org*, 2024, popp.undp.org/taxonomy/term/5246. Accessed 11 Dec. 2024.

Statista. “Artificial Intelligence - Global | Statista Market Forecast.” *Statista*, Aug. 2024, www.statista.com/outlook/tmo/artificial-intelligence/worldwide.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems Background, Mission and Activities of the IEEE Global Initiative. standards.ieee.org/wp-content/uploads/import/documents/other/ec_about_us.pdf. Accessed 14 Dec. 2024.

UNESCO. “Ethics of Artificial Intelligence.” *Www.unesco.org*, UNESCO, 2024, www.unesco.org/en/artificial-intelligence/recommendation-ethics.

Zeng, Yi. “Safe and Ethical AI (SEA) Platform Network · Linking Artificial Intelligence Principles (LAIP).” *Linking-Ai-Principles.org*, 2019, www.linking-ai-principles.org/principles. Accessed 14 Dec. 2024.