Protecting Patients' Human Rights from Algorithmic Biases in Healthcare: A Novel Ethical Governance Approach

Introduction

The IDx-DR, the first-ever artificial intelligence (AI) diagnostic tool authorized by the US Food and Drug Administration (FDA) in the medical field, significantly outperformed human clinicians in detecting eye diseases (Abràmoff et al., 2018). Innovative technology such as AI has advanced capabilities to help healthcare workers (HCWs) with "clinical data interpretation, clinical trial participation, image-based diagnosis, preliminary diagnosis, virtual nursing, and connected healthcare devices" (Asan et al., 2020). AI excels in such tasks because of its ability to learn from large datasets efficiently and perform tasks consistently without burnout. Therefore, AI systems have the potential to become essential tools to assist in medical decision-making within healthcare.

However, AI usage in healthcare raises severe ethical concerns, as they may threaten human rights to non-discrimination and equity (Geneva International Centre for Justice [GICJ], 2024). Many are worried about potential biases in AI training datasets due to insufficient information on marginalized populations and clinician biases (Asan et al., 2020). Reliance on biased AI systems may lead to negative consequences, as exemplified by Dr. Joy Buolamwini and Dr. Timnit Gebru's "Gender Shades" study. It revealed that current facial recognition algorithms were "less accurate in identifying darker-skinned individuals, particularly women," due to "technical flaws" and "the underrepresentation of diverse faces in the training datasets" (Huang et al., 2024). This example demonstrates that algorithmic bias could perpetuate colorism towards darker-skinned people.

This problem is more pronounced when AI's algorithmic biases are widely adopted. For instance, a health-risk-prediction algorithm that applied to nearly 200 million Americans annually was racially biased insofar that it calculated that "Black patients [were] considerably sicker than white

patients" (Obermeyer et al., 2019). As demonstrated, algorithmic bias significantly challenges the human rights of non-discrimination and equity, highlighting the need for solutions to address these ethical issues while utilizing technological innovation to benefit global populations.

In this paper, I will propose a human-centered AI (HCAI) governance structure inspired by the Total Product Lifecycle (TPLC) framework to effectively identify sources of and mitigate algorithmic bias to protect the integrity of human rights without stifling technological innovation. Furthermore, the IDx-DR case study will prove how strictly governed AI can enhance safety and equity in healthcare and beyond.

UNESCO Ethical Principles

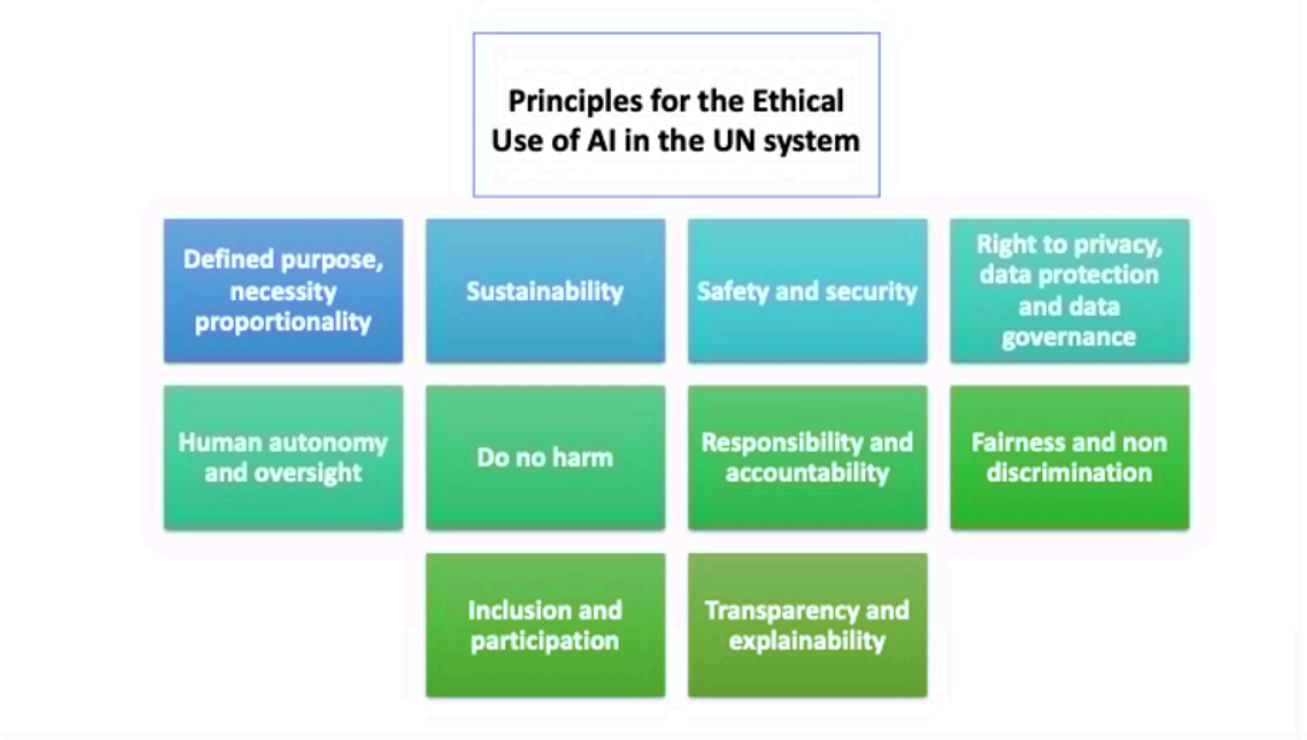
In the initial stage of addressing ethical issues, AI developers and HCWs should be aware of the public's concerns and needs. Asan et al. (2020) proposed three main pursuits to follow when designing AI:

1) Transparency: ensures secured data access and traceability to expose hidden bias
2) Robustness or reliable output: derives from high-quality and accurate data with low bias.
3) Fairness: enforces equitable outcomes across all demographics, allowing for widespread industry adoption.

All pursuits are related to reducing algorithmic bias. According to these goals, social organizations may establish additional ethical principles to encourage AI developers and HCWs to consider human rights and minimize algorithmic bias. UNESCO has proposed 10 principles for the ethical use of artificial intelligence in the United Nations Systems (GICJ, 2024).

Figure 1

Principles for the Ethical Use of AI in the UN system



Note: Shown in green and blue boxes, 10 principles are: (1) Defined purpose, necessity proportionality; (2) Sustainability; (3) Safety and security; (4) Right to privacy, data protection and data governance; (5) Human autonomy and oversight; (6) Do no harm; (7) Responsibility and accountability; (8) Fairness and nondiscrimination; (9) Inclusion and participation; (10) Transparency and explainability.

These principles serve as a critical framework to highlight ethical considerations in AI usage—this is especially important in life-endangering situations in healthcare. With such goals and principles established beforehand, AI developers and HCWs are encouraged to identify and eradicate algorithmic bias and are expected to uphold human rights.

Detecting Sources of Bias with the TPLC Model

To put UNESCO's ethical principles into practice, algorithmic bias must be identified at every stage of AI implementation. This may sound challenging, as sources of bias vary significantly. However, the TPLC, a proposed framework by computer engineer and healthcare specialist Michael Abràmoff et

al. (2023-a), provides significant insights into the process of detecting potential biases in AI models. The TPLC contains six phases and introduces potential biases in each phase:

(1) Conceptualization: bias stemming from exclusion of marginalized communities in historical datasets and differences in disease severity;

(2) Design: bias related to algorithm design due to a lack of ethical and clinical restraints;

(3) Development: bias related to training sets and reference standards;

(4) Validation: bias related to validation studies, human factors, and workflows;

(5) Access: affordability bias;

(6) Monitoring: bias caused by lack of ethical metrics;

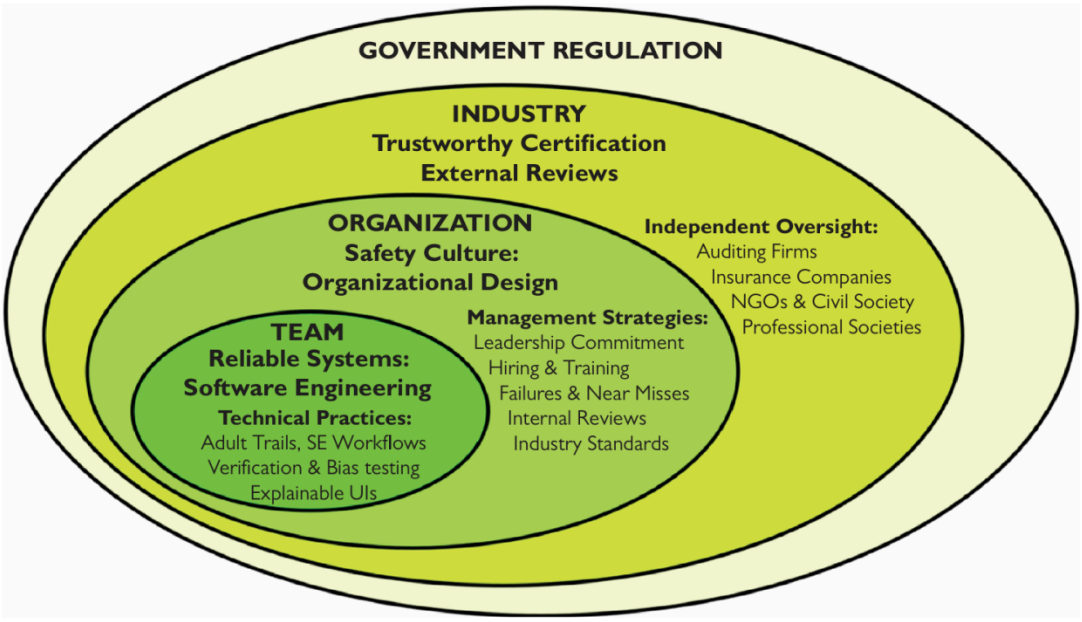
The TPLC provides a comprehensive guideline for HCWs, indicating how to appropriately implement AI. Crucially, Abràmoff et al. (2023-a) note that resolving biases in former phases does not necessarily prevent bias in latter stages, which tends to be overlooked by AI developers and HCWs. Therefore, the TPLC's comprehensiveness is highly valuable in presenting various biases for all parties to consider when adopting AI algorithms in healthcare, thereby protecting human rights.

Human-Centered AI Governance Structure

Under the TPLC, all involved parties are encouraged to collaborate to mitigate algorithmic bias in each phase of AI utilization. Ben Shneiderman (2022), author of the book "Human-Centered AI", introduced an HCAI governance structure that can be used in conjunction with the TPLC to guide related parties in conducting practical actions, as demonstrated below:

Figure 2

Governance Structures for Human-Centered AI



Note: The four-level governance structure is shown as green nested ovals. From the inner to the outer layer, the ovals present: (1) Team: reliable systems on software engineering; (2) Organization: safety culture through organizational design; (3) Industry: trustworthy certification through external reviews; (4) Government regulation.

In the conceptualization, design, and development phases of the TPLC, AI developers should train AI systems with diverse datasets to increase accuracy and reduce bias in its outputs. Additionally, well-designed AI systems should have explainable user interfaces, presenting the algorithms used and relevant scientific evidence supporting their final decision to HCWs and patients. Developers can achieve this by drafting safety cases—documents offering reasoning and evidence for decisions reached by AI (Habli et al., 2020). However, even with extensive planning, it is impossible to release perfected software. Thus, consistent evaluations and testing based on clinician feedback and accident reports are needed to consistently improve software engineering post-release, as emphasized by the US National Security Commission on AI (Schmidt et al., 2019). HCWs may conduct testing by referencing historical studies and scientific evidence to compare expected results with actual outputs.

AI industry leaders can contribute to the testing process by demanding developers and HCWs commit to responsible AI usage within all TPLC phases. Effective approaches include encouraging extensive reporting of failures. One example is the detailed testing of robotic surgery systems from the Manufacturer and User Facility Device Experience (MAUDE) database, examining 10,624 reports and discovering 8,061 device malfunctions (Alemzadeh et al., 2016). In this way, HCWs can discuss and reflect on past failures to gain insights into further improvement, thereby constructing an environment that pursues low biases and values human rights.

In the latter TPLC stages, social organizations could provide independent oversight. This could be achieved by planning oversight, continuous monitoring, and retrospective analysis of disasters (Shneiderman, 2022). The approaches together form a comprehensive structure for ensuring proper usage of AI and increasing pressure on technology companies designing AI, thus ensuring accurate diagnosis and enhancing trust. An inspiring example of such oversight is the Algorithmic Justice League. Under their independent examination, large technology companies improved their facial recognition products by reducing gender and racial bias within two years (Shneiderman, 2022).

Government regulations further promote effective monitoring to reduce algorithmic bias and ensure the secure usage of AI. Governments can fund AI research, encouraging experts to study specific cases and suggest AI policy implications (Shneiderman, 2022). The US Congress has demonstrated this by funding the US National Transportation Safety Board (NTSB) to send teams of experts to conduct field research on AI incidents. This allowed the NTSB to provide insightful reports on current AI limitations and impose pressure on AI developers and HCWs to reduce algorithmic biases. Shneiderman (2022) further acknowledged the approach of the FDA, focusing on AI systems in medical fields and designing outlines for careful regulation. Specifically, it encourages discussions on how AI may improve the performance of current medical devices; it also takes responsibility for confirming the accuracy and

efficiency of AI systems. Favorable AI performances correspond to low algorithmic bias. Therefore, the FDA reduces algorithmic bias and enhances patient trust in the AI systems it authorizes and regulates, protecting their human rights to non-discrimination and equity.

Deploying the governance structure in alignment with proposed ethical frameworks and the TPLC model has proven effective through successful AI system implementations. The IDx-DR system—a fully autonomous AI solution for detecting diabetic eye disease—diabetic retinopathy (DR) has been tested to be effective and is currently implemented in multiple healthcare centers (Abràmoff et al., 2018). The IDx-DR was designed carefully with ethical considerations aligning with the TPLC and HCAI governance structure.

When developing DR indicators, software engineers referred to decades of research regarding its diagnosis and management to address the historical access disparities, reducing biases in the TPLC conceptualization phase. They utilized various DR indicators as biomarkers for the AI system to process before reaching its conclusion. This is important, as "using multiple, statistically dependent detectors for such lesions, each optimized using machine learning algorithms, addresses equity in the design phase"(Abràmoff et al., 2020). As a biomarker-based AI system, IDx-DR further excels by providing indications and explanations for its conclusion, reducing incorrect diagnoses and bias due to the inability of human clinicians to validate AI results. In this way, they established sound AI software engineering.

Researchers used multiple metrics to measure the accuracy and potential bias in the generated outcomes. Sensitivity determines the proportion of correctly detected positives in DR; specificity determines the proportion of correctly detected negatives in DR. Overall, high accuracy is reflected through high sensitivity and specificity values. Equity was analyzed through the diagnosability metric, measuring the number of patients who received valid rather than indeterminate results. If only a few people could be fully diagnosed within a sample representing their population, then researchers would

conclude that the AI failed to guarantee equity (Abràmoff et al., 2020). Former studies have shown the sensitivity, specificity, and diagnosability rates for experienced clinicians were 30-40%, 95%, and 80-90%, respectively. IDx-DR, on the other hand, presented results of 87%, 91%, and 96% for the three metrics, respectively, exceeding the thresholds of clinical experts in detecting AI (Abràmoff et al., 2018). Such validation testing by AI developers and individual researchers enhances high diagnosis accuracy and protects human rights. Given its high accuracy and ethical considerations to protect equity throughout development, the FDA authorized it for use by HCWs (Abràmoff et al., 2023-b). This would also require the FDA to consistently inspect IDx-DR in the future, monitoring for potential biases in the latter TPLC stages.

Conclusion

The example of IDx-DR confirms the effectiveness of putting the HCAI governance structure into practice in health centers following ethical principles and frameworks such as the TPLC to reduce bias. Healthcare is one of the most complex fields for AI implementation, as patient health is directly influenced, and negative outcomes can be life-critical. Therefore, this solution has a high potential to work in other areas that utilize AI to protect human rights. However, approaches to achieving the governance structure must be comprehensively discussed among experts before implementation. All stakeholders should consistently protect human rights when implementing AI to achieve sustainable and ethical development.

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Nature Partner Journals Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0040-6

Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., Eydelman, M. B., & Maisel, W. H. (2023-a). Considerations for addressing bias in artificial intelligence for health equity. *Nature Partner Journals Digital Medicine*, *6*(1), 1–7. https://doi.org/10.1038/s41746-023-00913-9

Abràmoff, M. D., Tobey, D., & Char, D. S. (2020). Lessons learned about autonomous AI: Finding a safe, efficacious, and ethical path through the development process. *American Journal of Ophthalmology*, *214*, 134–142. https://doi.org/10.1016/j.ajo.2020.02.022

Abràmoff, M. D., Whitestone, N., Patnaik, J. L., Rich, E., Ahmed, M., Husain, L., Hassan, M. Y., Tanjil, M. S. H., Weitzman, D., Dai, T., Wagner, B. D., Cherwek, D. H., Congdon, N., & Islam, K. (2023-b). Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. *Nature Partner Journals Digital Medicine*, *6*(1), 1–8. https://doi.org/10.1038/s41746-023-00931-7

Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z., & Iyer, R. K. (2016). Adverse events in robotic surgery: A retrospective study of 14 years of FDA data. *PLOS One*, *11*(4), e0151470. https://doi.org/10.1371/journal.pone.0151470

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, *22*(6). https://doi.org/10.2196/15154

Geneva International Centre for Justice [GICJ]. (2024, June 17). *Digital rights: The impact of AI and emerging technologies on human rights*. https://www.gicj.org/topics/thematic-issues/business-human-rights/3758-digital-rights-the-impact-of-ai-and-emerging-technologies-on-human-rights

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in healthcare: Accountability and safety. *Bulletin of the World Health Organization*, *98*(4), 251–256. https://doi.org/10.2471/blt.19.237487

Huang, J., Lane, C., & Manning, P. (2024). *The impact of bias in AI-driven healthcare: Challenges and considerations for equitable implementation*. OxJournal. https://www.oxjournal.org/the-impact-of-bias-in-ai-driven-healthcare/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. https://www.ftc.gov/system/files/documents/public_events/1548288/privacycon-2020-ziad_obermeyer.pdf

Schmidt, E., Work, R. O., Catz, S., Chien, S., Clyburn, M. L., Darby, C., Ford, K., Griffiths, J. M., Horvitz, E., Jassy, A., Louie, G., Mark, W., Matheny, J., Mcfarland, K., & Moore, A. W. (2019). *National Security Commission on Artificial Intelligence: Interim report, November 2019.* UNT Digital Library. https://digital.library.unt.edu/ark:/67531/metadc1851191/

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press. https://doi.org/10.1093/oso/9780192845290.001.0001