Towards AI algorithmic fairness

Khensani Xivuri: PhD candidate in IT Management

University of Johannesburg

Introduction

Despite Artificial Intelligence (AI) being one of the fastest-growing fields due to its ability to enhance competitive advantage in organisations, there are growing concerns around its inherent bias.

AI allows systems to perform tasks that would normally be performed by humans. Data is collected and processed to provide results that mimic human intelligence over rules learnt over time. There are three categories of AI, namely, Narrow AI, General AI and Super AI, and some of the popular AI's are natural language processing, computer vision, and artificial neural networks. Apple's Siri, Google Maps, Google predictions and Smart replies by GMAIL are some of the well-known AI solutions.

Fairness means impartiality in decision-making, making decisions without any self-interest in the outcome. AI is prone to AI algorithmic bias, which could result in operational and reputational damage. AI has the potential to be unfair to certain groups of people like race and gender. AI systems need to be fair and respect human rights, democracy, values, and principles.

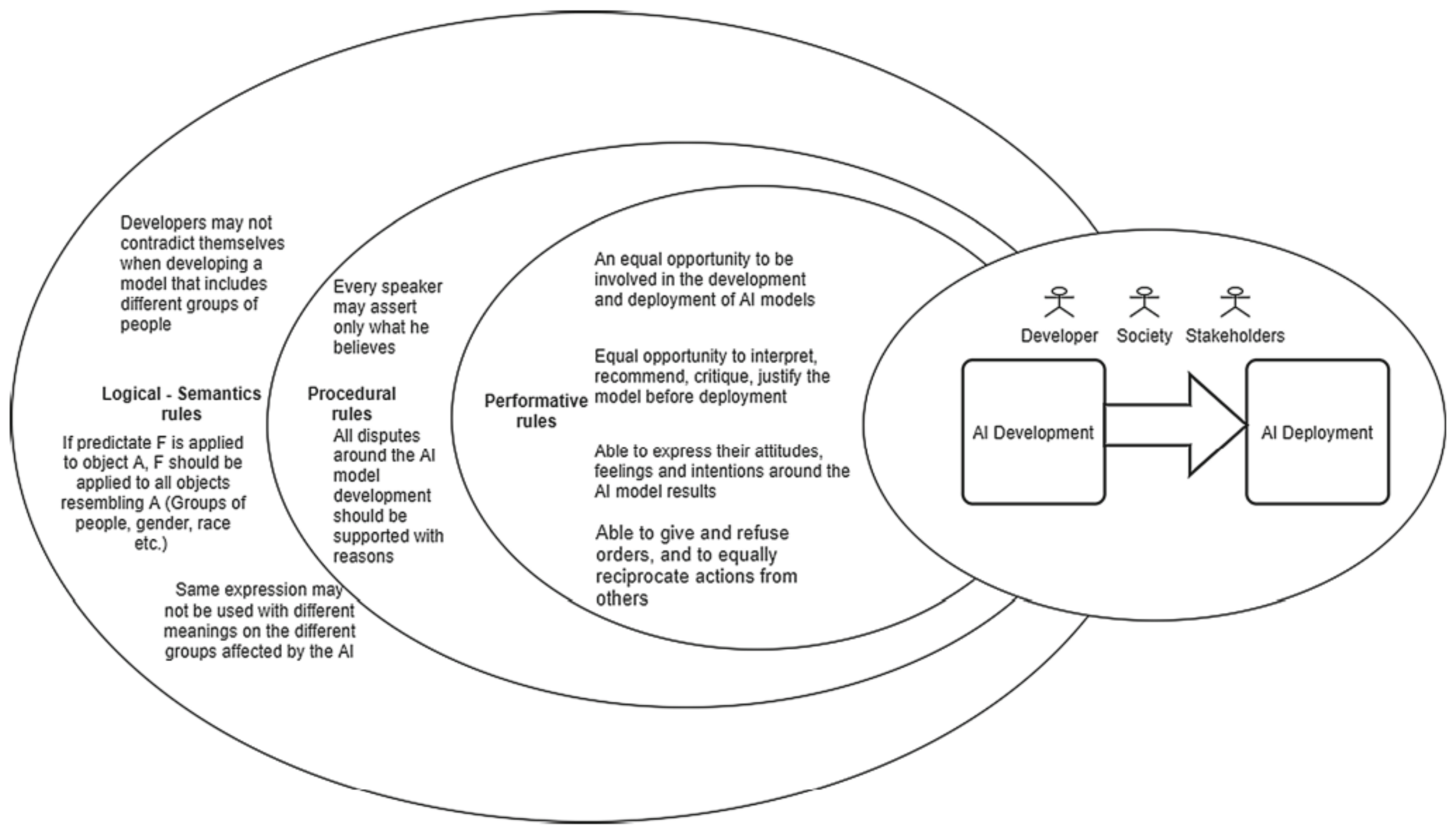
AI development is always focused on the results and not the input and bias in the data or process of developing the AI. This has resulted in biased AI being implemented, and such issues only picked up by the public. A framework that supports the emphasis on shifting the focus from AI standards to **fair processes** that are adaptable to AI's continuous innovation has been developed.

The Jurgen Habermasian critical theory of communicative action, the lifeworld and meaning was used to develop a process framework on AI algorithmic fairness.

❑ The framework engages logical-semantic, procedural and performative rules that can be applied to the development process of AI to avoid any domination before, during and after the AI development process.

Moral arguments
Society free of domination
Justification of arguments/views
How people experience reality
The life world
The life world
Moral arguments
Society free of domination
People and their perceptions
Common understanding
The life world

The purpose of this research is to collect data and demonstrate the Habermasian approach to AI Fairness processes framework.


Developers may not contradict themselves when developing a model that includes different groups of people
Logical - Semantics rules
If predictate F is applied to object A, F should be applied to all objects resembling A (Groups of people, gender, race etc.)
Same expression may not be used with different meanings on the different groups affected by the AI
Every speaker may assert only what he believes
Procedural rules
All disputes around the AI model development should be supported with reasons
Performative rules
An equal opportunity to be involved in the development and deployment of AI models
Equal opportunity to interpret, recommend, critique, justify the model before deployment
Able to express their attitudes, feelings and intentions around the AI model results
Able to give and refuse orders, and to equally reciprocate actions from others
Developer
Society
Stakeholders
AI Development
AI Deployment

Methods

The **critical realism** philosophical approach will be used to explain the outcomes of the framework that will be demonstrated during this research. Critical research focuses on domination, power relations, conflicts and contradictions in society. Critical realism will be used to show why society should be involved throughout the different stages of AI development. The research will demonstrate the importance of following the Habermas discursive requirements throughout the development phases of AI.

This research focuses on reducing bias in AI algorithms and, therefore, will be using the **qualitative research design**. The qualitative research approach is best suited for research that involves social justice and communities. The qualitative research approach examines the meaning that individuals or specific groups ascribe to, using research questions and data to answer the research question.

Design science research will be used to develop a solution that will use the developed framework to reduce bias in AI algorithms. The solution will allow for the different parties to be involved in the whole process of AI development. Professionals in the AI development field can use the solution to reduce bias in AI algorithms. The solution developed will be used by developers, the society that will be affected by the AI and management sponsoring the development of the AI.

Population and Sample

The Habermasian approach to fair processes in AI will be presented to AI professionals for their feedback. AI professionals will see the value in using such a framework when developing AI.

A solution for reducing AI bias in algorithms will be built based on the framework. The solution will be used by developers, society and management, demonstrating the framework during an AI development project. The built solution will be used to capture all inputs from the different parties.

❑ **Management** will use the solution to capture their requirements and expectations of the AI.
❑ **Developers** will use the app to capture the details of the development approach, input data and rules that will be used in the development.
❑ **Society** will use the app to question any areas of concern and express their feelings.

All parties involved will use the solution for any disputes and reasons for their disputes.

Any bias will be picked up during the development stages and resolved before the AI solution is implemented in the live environment.

Expected Contributions

❑ Involving affected parties, stakeholders and developers in the development and implementation stages of AI, and incorporating the logical-semantic, procedural, and performative rules during these stages is an important first step.

❑ Using the solution will help ensure that all parties are involved throughout the different stages of AI implementation.

❑ Any bias will be picked up before the application is implemented onto the live environment.

❑ Any bias picked up before implementation will be resolved before implementation, avoiding any backlash from society, and any reputational damage or financial loss to the organisation.

Publications

❑ A Systematic review of AI algorithmic Fairness - Xivuri, K. and Twinomurinzi, H. (2021) *A Systematic Review of Fairness in Artificial Intelligence Algorithms, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer International Publishing. doi: 10.1007/978-3-030-85447-8_24.

❑ A Habermasian approach to fair Processes in AI algorithms – Xivuri, K. and Twinomurinzi, H. (2022) 'A Habermasian Approach to Fair Processes in AI Algorithms', *SACAIR 2021*, 1.

Bibliography

1. Creswell, W. J. and Creswell, J. D. (2018) *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Fifth Edit, *SAGE Publications, Inc.* Fifth Edit. SAGE Publications. Available at: file:///C:/Users/Harrison/Downloads/John W. Creswell & J. David Creswell - Research Design_ Qualitative, Quantitative, and Mixed Methods Approaches (2018).pdf%0Afile:///C:/Users/Harrison/AppData/Local/Mendeley Ltd./Mendeley Desktop/Downloaded/Creswell, Cr.
2. Farnadi, G., Babaki, B. and Getoor, L. (2018) 'Fairness in Relational Domains', *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 108–114. doi: 10.1145/3278721.3278733.
3. Ghosh, A., Chakraborty, D. and Law, A. (2018) 'Artificial intelligence in Internet of things', *CAAI Transactions on Intelligence Technology*, 3(4), pp. 208–218. doi: 10.1049/trit.2018.1008.
4. Habermas, J. (1984) *THE THEORY OF COMMUNICATIVE ACTION*.
5. Habermas, J. (1990) *Moral Consciousness and Communicative Action: Moral Conciousness and Communicative Action (Studies in Contemporary German Social Thought)*.
6. IEEE (2017) *New IEEE Standards for Artificial Intelligence Affecting Human Well-Being*. Available at: https://transmitter.ieee.org/new-ieee-standards-artificial-intelligence-affecting-human-well/ (Accessed: 28 August 2021).
7. Neuteleers, S., Mulder, M. and Hindriks, F. (2017) 'Assessing fairness of dynamic grid tariffs', *Energy Policy*. Elsevier Ltd, 108(September), pp. 111–120. doi: 10.1016/j.enpol.2017.05.028.
8. Oates, B. (2006) *Researching Information Systems and Computing*.
9. Salah, K. *et al.* (2019) 'Blockchain for AI: Review and open research challenges', *IEEE Access*. IEEE, 7, pp. 10127–10149. doi: 10.1109/ACCESS.2018.2890507.
10. Scherer, M. U. (2016) 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies', *Havard Journal of Law & Technology*. Elsevier BV, 29(2), pp. 354–200. doi: 10.2139/ssrn.2609777.