Multi-Objective Fairness Approach Using Causal Bayesian Networks & Grammatical Evolution

Zahid Irfan*
zahid.irfan@dkit.ie
Dundalk Institute of Technology
Dundalk, Ireland

Róisín Loughran
roisin.loughran@dkit.ie
Dundalk Institute of Technology
Dundalk, Ireland

Muhammad Adil Raja
adil.raja@dkit.ie
Dundalk Institute of Technology
Dundalk, Ireland

Fergal McCaffery
fergal.mccaffery@dkit.ie
Dundalk Institute of Technology
Dundalk, Ireland

Abstract

Addressing unwanted biases has become critical as Artificial Intelligence systems are increasingly integrated into various aspects of society. Bias in decision-making can lead to unfair outcomes, perpetuating social inequalities and discrimination. Causal graphs enable the identification of causal mechanisms that may contribute to biased outcomes. Evolutionary computation techniques are well known for exploring large, complex solution spaces and evolving optimal solutions over successive generations. We propose a novel approach that combines causal structures with grammatical evolution, a method using grammar, to create directed acyclic graphs for modelling and evolving solutions using fairness and accuracy as fitness criteria. Our approach evolves causal graphs that balance model fairness and performance in single-objective and multi-objective settings. Results show that the multi-objective optimization improved fairness by 32 percent while reducing accuracy by only 2.85 percent compared to the single-objective case. This demonstrates that integrating causal mechanisms with evolutionary computation can effectively develop Artificial Intelligence systems that are both accurate and fair.

Keywords

Artificial Intelligence, Machine Learning, Fairness, Bias, Causal Models, Grammatical Evolution, Causal Bayesian Networks

ACM Reference Format:
Zahid Irfan, Róisín Loughran, Muhammad Adil Raja, and Fergal McCaffery. 2025. Multi-Objective Fairness Approach Using Causal Bayesian Networks & Grammatical Evolution . In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Corresponding author.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Unwanted bias refers to the unconscious assumptions that affect our decisions, actions, and interactions [6]. A pervasive issue in various aspects of life, bias has been studied in areas of philosophy and psychology [1, 14]. Frequently these biases are present due to societal and behavioural factors, like culture, personal experiences, and environmental influences such as stereotyping, confirmation bias, authority bias, and overestimation bias. Sometimes bias can be present due to involuntary factors including, anchoring, recall bias, recency bias, halo effect, or attribution bias [9]. While not all biases lead to negative outcomes, certain unwanted biases can perpetuate discriminatory behaviours or actions, resulting in unfair treatment and manifesting cycles of inequality and injustice [8]. It is essential to recognize and address these unwanted biases to prevent unintended consequences and promote a more equitable and just society. Protected attributes such as race, gender, or age are legally protected against discrimination [5]. Unprivileged groups face potential disadvantage compared to privileged groups in contexts like hiring, and financial lending [12].

Artificial Intelligence (AI) systems were built to assist decision-making [4]. Unwanted bias can have harmful consequences when AI systems are applied to real world problems, since it can result in social inequalities and discrimination. Strategies to detect, mitigate, and prevent unwanted bias in AI systems are essential to ensure fair, transparent and unbiased systems. There have been concerns about bias in AI systems since the earliest days of its development [10, 11]. The outcomes of an AI system and any decision made by such a system were dependent on the data and could have been subject to unwanted bias[12]. Facial recognition systems discriminated against people labelled female, Black, or between the ages of 18—30 than for other demographic cohorts [3]. Bias was also detected in other systems like decision support systems in criminal justice systems, healthcare, financial systems [5]. Recognizing bias in AI systems has prompted a growing body of research and public discourse on understanding its causes, manifestations, and consequences leading to development of strategies to mitigate its impact [2]. Fairness metrics are used to determine the extent of the predictions of an AI system are biasing the results in favour of some individual or group [7].

In this paper, we explore how Evolutionary Computation (EC) techniques can be utilized to harness causal relationships among features in a dataset, aiming to develop models that optimise the key aspects of AI systems, specifically targeting improvements in fairness and accuracy. Causal theory, as pioneered by Judea Pearl, is a fundamental framework for analysing dependencies and understanding causal relationships in complex systems [13]. We will be evolving causal graphs using Grammatical Evolution (GE) to build Bayesian Networks (BNs) for optimising fairness and accuracy.

The rest of the paper is organized as the following. The related work is discussed in Section **??**, followed by methodology in Section **??**. We present our results in Section **??** which are discussed in Section **??**. Finally, we offer some conclusions and propose the future directions in Section **??**.

2 Introduction

Unwanted bias refers to the unconscious assumptions that affect our decisions, actions, and interactions [6]. A pervasive issue in various aspects of life, bias has been studied in areas of philosophy and psychology [1, 14]. Frequently these biases are present due to societal and behavioural factors, like culture, personal experiences, and environmental influences such as stereotyping, confirmation bias, authority bias, and overestimation bias. Sometimes bias can be present due to involuntary factors including, anchoring, recall bias, recency bias, halo effect, or attribution bias [9]. While not all biases lead to negative outcomes, certain unwanted biases can perpetuate discriminatory behaviours or actions, resulting in unfair treatment and manifesting cycles of inequality and injustice [8]. It is essential to recognize and address these unwanted biases to prevent unintended consequences and promote a more equitable and just society. Protected attributes such as race, gender, or age are legally protected against discrimination [5]. Unprivileged groups face potential disadvantage compared to privileged groups in contexts like hiring, and financial lending [12].

AI systems were built to assist decision-making [4]. Unwanted bias can have harmful consequences when AI systems are applied to real world problems, since it can result in social inequalities and discrimination. Strategies to detect, mitigate, and prevent unwanted bias in AI systems are essential to ensure fair, transparent and unbiased systems. There have been concerns about bias in AI systems since the earliest days of its development [10, 11]. The outcomes of an AI system and any decision made by such a system were dependent on the data and could have been subject to unwanted bias[12]. Facial recognition systems discriminated against people labelled female, Black, or between the ages of 18—30 than for other demographic cohorts [3]. Bias was also detected in other systems like decision support systems in criminal justice systems, healthcare, financial systems [5]. Recognizing bias in AI systems has prompted a growing body of research and public discourse on understanding its causes, manifestations, and consequences leading to development of strategies to mitigate its impact [2]. Fairness metrics are used to determine the extent of the predictions of an AI system are biasing the results in favour of some individual or group [7].

In this paper, we explore how EC techniques can be utilized to harness causal relationships among features in a dataset, aiming to develop models that optimise the key aspects of AI systems, specifically targeting improvements in fairness and accuracy. Causal theory, as pioneered by Judea Pearl, is a fundamental framework for analysing dependencies and understanding causal relationships in complex systems [13]. We will be evolving causal graphs using GE to build BNs for optimising fairness and accuracy.

The rest of the paper is organized as the following. The related work is discussed in Section **??**, followed by methodology in Section **??**. We present our results in Section **??** which are discussed in Section **??**. Finally, we offer some conclusions and propose the future directions in Section **??**.

Acknowledgments

This research was supported through the HEA's Technological University Transfer Fund (TUTF) and Dundalk Institute of Technology (DkIT).

References

[1] Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare* 9 (2016), 211–217. doi:10.2147/JMDH.S104807 arXiv:https://www.tandfonline.com/doi/pdf/10.2147/JMDH.S104807 PMID: 27217764.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

[3] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

[4] Daniel Castro and Joshua New. 2016. The promise of artificial intelligence. *Center for data innovation* 115, 10 (2016), 32–35.

[5] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 339–348. doi:10.1145/3287560.3287594

[6] Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics* 18 (2017), 1–18.

[7] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems* 2 (2016), 3315–3323.

[8] D. Kahneman, P. Slovic, and A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press. https://books.google.ie/books?id=_0H8gwj4a1MC

[9] Daniel Kahneman, Amos Tversky, et al. 1977. *Intuituve Prediction: Biases and Corrective Procedures*. Decision Research, Perceptronics.

[10] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 3819–3828.

[11] Kristin M Kostick-Quenet, I Glenn Cohen, Sara Gerke, Bernard Lo, James Antaki, Faezah Movahedi, Hasna Njah, Lauren Schoen, Jerry E Estep, and JS Blumenthal-Barby. 2022. Mitigating racial bias in machine learning. *Journal of Law, Medicine & Ethics* 50, 1 (2022), 92–100.

[12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[13] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (12 1995), 669–688. doi:10.1093/biomet/82.4.669 arXiv:https://academic.oup.com/biomet/article-pdf/82/4/669/698263/82-4-669.pdf

[14] Carolyn Wood Sherif. 1998. Bias in psychology. *Feminism & Psychology* 8, 1 (1998), 58–75.