International Journal of Law and Information Technology, 2021, 00, 1–16
doi: 10.1093/ijlit/eaab008
Article

OXFORD

AI ethical bias: a case for AI vigilantism (AIlantism) in shaping the regulation of AI

Ifeoma Elizabeth Nwafor*

ABSTRACT

The debate on the ethical challenges of artificial intelligence (AI) is nothing new. Researchers and commentators have highlighted the deficiencies of AI technology regarding visible minorities, women, youth, seniors and indigenous people. Currently, there are several ethical guidelines and recommendations for AI. These guidelines provide ethical principles and humancentred values to guide the creation of responsible AI. Since these guidelines are non-binding, it has no significant effect. It is time to harness initiatives to regulate AI globally and incorporate human rights and ethical standards in AI creation. The government need to intervene, and discriminated groups should lend their voice to shape AI regulation to suit their circumstances. This study highlights the discriminatory and technological risks suffered by minority/marginalised groups owing to AI's ethical dilemma. As a result, it recommends the guarded deployment of AI vigilantism to regulate the use of AI technologies and prevent harm arising from AI systems' operations. The appointed AI vigilantes will comprise mainly persons/groups with an increased risk of their rights being disproportionately impacted by AI. It is a well-intentioned group that will work with the government to avoid abuse of powers.

KEYWORDS: Artificial Intelligence, AI vigilantism, Marginalised groups inclusion, AI ethical bias, AI discrimination, Regulating AI, Global AI governance, Vigilantes

INTRODUCTION

In recent years the discourse on artificial intelligence (AI) is perceived with both enthusiasm and fear.[1] AI has numerous benefits like boosting the economy by enhancing the evolution of work and ensuring efficiency in sectors where it is employed. However, the ethical challenges surrounding AI and the absence of diversity in the industry are some of the areas of enormous concern. It has steered to unfair and illicit discrimination.[2]

* Ifeoma Elizabeth Nwafor, Lecturer, Faculty of Law, Godfrey Okoye University, Enugu State, Nigeria. Email: ifeomanwafor900@gmail.com

1 Mark Coeckelbergh, *AI Ethics* (The MIT Press 2020).

2 Frederick Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73> last accessed 21 September 2021.

Most questions relating to the debate on AI's ethics and governance are centred on addressing the misinformation on AI,[3] identifying the ethical issues caused by AI, and examining the part government has played in fashioning AI ethical standards.[4]

Gordon et al.[5] comprehensively explored the impact of AI ethical issues on human society. They examined the ethical relevance of AI by highlighting the importance of implementing ethics for AI systems which will enhance safety guidelines and prevent risks for humanity. The article concentrates on debates that have received attention in the AI field. Debates such as machine ethics and bias, opacity issues, the concern about human enfeeblement and anthropomorphism. The authors briefly analysed various AI ethical guidelines and highlighted the need to include African and Asian perspectives in drafting such policies.

Resseguier and Rodrigues[6] underscore the nature of various AI ethics developed in response to the adverse impact of AI on human society. They stressed that the principled approach to AI ethics is problematic and limits its effectiveness. This problem stems from the misuse of ethics in replacement for regulation and not properly harnessing the ethical teeth in the field. They argued that the law conception of ethics, dominant in current AI ethics, is not ethics itself but rather an end of ethics.

Despite the various articles on AI ethical bias, gaps remain on the discriminatory risks suffered by marginalized groups and the role such groups can play to mitigate such risks. This is where this article comes in. The question, whether AI risks and discrimination can be mitigated by the inclusion of disadvantaged groups at the design and regulation-making stage has been accorded little attention. In considering

3 Deborah G Johnson and Mario Verdicchio, 'Reframing AI Discourse' (2017) <https://link.springer.com/article/10.1007/s11023-017-9417-6> last accessed 13 August 2021. The authors identified AI research and AI products as a fundamental ethical issue facing the AI research community. They added that the public's understanding and confusion on AI might hinder AI research. The authors recommend the creation and implementation of language to reconstruct the debate on AI and highlight the actual issue in the field. James Vincent, 'What Counts as Artificial Intelligence? AI and Deep Learning, Explained' [2016] <https://www.theverge.com/2016/2/29/11133682/deep-learning-ai-explained-machine-learning> last accessed 8 July 2021. The author details the difficulty in defining and using the term artificial intelligence. Ralf T Kreutzer and Marie Sirrenberg, 'What is Artificial Intelligence and How to Exploit it? In: Understanding Artificial Intelligence' Management for Professionals Springer (2020). <https://doi.org/10.1007/978-3-030-25271-7_1> last accessed 19 September 2021. The authors provides definitions of crucial terms surrounding AI and highlight the relationships of terms such as machine learning and deep learing which is mostly perceived as AI by the public. The authors also identify the global economic impact on AI.

4 Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019). The author explored the current trends in government AI regulation from the UK, France, European Union, USA, Japan and China. He notes that national AI regulations are immersed with the nations' positions in the global rank. He opines that China's effort in shaping global AI regulation has political and economic undertone. 'Artifical Intelligence: From Ethics to Policy' [June 2020] <https://www.europarl.europa.eu> last accessed 13 August 2021. Olivia J Erdelyi Judy Goldsmith, 'Regulating Artificial Intelligence: Proposal for a Global Solution' (2018) <https://doi.org/10.1145/3278721.3278731> last accessed 13 September 2021. Ali Hahmi, 'AI Ethics: The Next Big Thing in Government' World Government Summit <https://www.worldgovernmentsummit.org> last accessed 13 August 2021.

5 John-Stewart Gordon and Sven Nyholm, 'Ethics of Artificial Intelligence/Internet Encyclopedia of Philosophy' (2021) <https://iep.utm.edu/ethic-ai/> last accessed 17 September2021.

6 Anais Resseguier and Rowena Rodrigues, 'AI Ethics should not Remain Toothless! A Call to Bring Back the Teeth of Ethics' (2020) 7 Big Data & Society DOI: 10.1177/2053951720942541 last accessed 19 September 2021.

this question, this article analyses the effect of the various ethical guidelines and highlights the need for government regulation. The impact of involving disadvantaged groups in the course of regulating AI has not been harnessed. Human impact plays a massive role in the creation and operation of AI models and tools. In the same breath, humans will be instrumental in preventing bias at the design, training, trial and final stage of AI tools. The inclusion of marginalized groups at the data gathering and design stage will hugely influence the outcome of such models. This article suggests the guarded deployment of AI vigilantism (AIlantism) to regulate the use of AI technologies and prevent harm arising from AI systems' operation. Vigilantes are perceived as groups seeking to achieve justice through unauthorized means. For instance, the actions of cyber-vigilantism[7] have been criticized for violating computer hacking laws. The study understands that cyber vigilantism is a grassroots initiative. It highlights that the positive impacts in the AI realm will outweigh its disadvantages. It proposes that such initiative in the AI field should be accepted legally by including members of minority/ marginalized groups to lend their voice to shape AI regulation to suit their circumstances. AI vigilantes will work closely with the government and law enforcement agencies to avoid abuse of statutory provisions. In recent times, there have been increased activism of indigenous communities developing guidelines on using their indigenous knowledge. This article explores this analogy concerning AI vigilantes lending their voice to shape AI regulation to suit their circumstances.

This article has six parts; part one covers the introduction; part two examines the risks and harms suffered by marginalized groups due to AI technology. While part three explores the impact of various AI ethical and legal guidelines/measures. Part four discusses various views on efforts to mitigate AI risks. Part five provides recommendations for promoting a responsible AI, and part six offers concluding remarks.

AI DISCRIMINATORY RISKS/HARMS

AI generates numerous discriminatory and technological harms. These risks include accidents, privacy violations, and life loss in cases where an AI medical algorithm goes wrong.[8] This article will focus on discriminatory risks suffered by individuals in the use of AI. Discrimination is 'treating differently, without an objective and reasonable justification, persons in analogous, or relevantly similar situations'.[9] AI can discriminate based on different grounds such as sex, gender identity, sexual orientation, race, ethnicity, age, disability, religion, language, nationality and social origin. Barocas and Selbst provide five different ways that can lead to unintentional discrimination.[10] These discriminations can arise from how 'class labels' are defined, the labelling or collection of training data, feature selection and the use of AI systems for intentional

7 Cyber vigilantes are private groups that combat cybercrime. For instance, Anonymous, an independent cyber vigilante group have recorded successes to combat cyber terrorism etcetera.

8 Benjamin Cheatham, Kia Javanmardian and Hamid Samandari, 'Confronting the Risks of Artificial Intelligence' (2019) <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence> last accessed 14 September 2021.

9 *Kiyutin v Russia* (App no 2700/10) ECHR 10.

10 Bryce T Daniel, 'Disparate Impact: A Failed Remedy for Discrimination' (2020) <https://dspace2.creighton.edu> last accessed 22 September 2021.

discriminatory ends.[11] AI systems tainted by historical biases may reflect these biases and disadvantage certain groups based on their history.[12] This article will focus on AI discrimination based on gender, sexual orientation/gender identity and race. It is taken seriatim below.

AI Discrimination Based on Gender

Over time the AI community has been presented as creating impartial and objective solutions to human problems. However, this is not a correct representation of the actual state of events as data solutions often carry bias and subjectivity mimicking its creators. For example, giant tech companies like Facebook or Google have less than 15 per cent of women working for them.[13] Above 80 per cent of AI professors are men, which means a more significant percentage of the feminine gender is unrepresented in the AI community. At the same time, little or no data exist on transgender workers.[14]

Generally, various statistical reports reveal that the AI Community is male-dominated; at a time, Nvidia only had 17 per cent employed female staff, Intel and Microsoft 26 per cent, Dell 28 per cent, and Google, Salesforce and YouTube 31 per cent. These reports often exclude transgender and how the diversity gap broadens at the most senior levels.[15] However, a survey report in 2020[16] reveals that females compose 28 per cent of the science and engineering workforce. Only 13.8 per cent of women have authored artificial intelligence-related research papers, and less than a quarter are considered AI professionals.[17] This statics reveals the rate at which gender discrimination permeates the AI community.

There is a link between AI and discrimination based on gender. Such discrimination is recorded in various fields such as machine translation tools, targeted advertising, and chatbox.[18] Conversational AI is an evolving computer interaction area that uses ordinary language to swap information and pass commands to computers.[19] Google translate is the most used language-translation tool, which translates one hundred and three languages. Researchers found that Google Translate converted gender-neutral pronouns to gender-stereotyped pronouns.[20] Research by Caliskan,

11 Borgesius (n 2).

12 Allen Smith, 'AI: Discrimatory Data In, Discrimination Out' (2019) <https://www.shrm.org/resourcesandtools/legal-and-compliance/employment-law/pages/artificial-intelligence-discriminatory-data.aspx> last accessed 19 September 2021.

13 ibid.

14 Kari Paul, 'Disastrous lack of Diversity in AI Industry Perpetuates Bias' *The Guardian* (16 April 2019) <https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack of diversity-new-york university-study> last accessed 20 August 2021.

15 James Stanie, 'We Must Fix AI's Diversity Problem' (2019) <https://medium.com/@jstanier/we-must-fix-ais-diversity-problem-6ad5fddc2f8c> last accessed 20 September 2021.

16 Paul (n 14).

17 <https://syncedreview.com/2020/03/13/exploring-gender-imbalance-in-ai-numbers-trends-and-discussions/> last accessed 15 August 2021.

18 Can Yvuz, 'Machine Bias Artificial Intelligence and Discrimination' (2019) <https://www.researchgate.net/publication/334721591_Machine_Bias_Artificial_Intelligence_and_Discrimination> last accessed 20 September 2021.

19 Urwa Muaz, 'Racila Bias in Conversational Artificial Intelligence' (2019) <https://towardsdatascience.com/racial-bias-in-conversational-artificial-intelligence> last accessed 19 September 2020.

20 Yvuz (n 18).

Bryson and Narayanan shows 'that machines can learn word associations from written texts and that these associations mirror those learned by humans'.[21] AI translation tools unavoidably reflect the biases contained in prejudiced natural languages.

In 2016, Microsoft released Tay, an AI chatbox designed to have a human-like conversation with Twitter users. Tay became racist, sexist, and anti-Semitic in less than 24 hours on Twitter.[22] Twitter chatbox developers argue that Microsoft failed to utilize a blacklist to restrain hate speech.[23] The blacklist is perceived as the technical solution to make AI racist talk-proof.

Facebook Inc has been accused of serving discriminatory advertisements based on race regarding AI discrimination and targeted advertising, resulting in the US Fair Housing Act's infringement.[24] Its Ad algorithm has a record of distributing discriminatory ads based on self-identified gender. The publication of evident ads indicating a preference based on race, sex, religion or other precise group is illegal under US law, such as the Federal Fair Housing Act.[25] One such case was when the Facebook algorithm, which excludes women from viewing online job ads. There are other algorithms with gendered inputs making sexist decisions. In 2013, Sweeney bared that Google displayed advertisements suggesting that a person had arrest records when African American sounding names were searched. Google systems inherited racial bias by evaluating people's surfing behaviour.[26] Only pictures of white babies appear when you type a cute baby on Google, which translates as discrimination against other descendants' babies.

AI Discrimination Based on Sexual Orientation/Gender Identity

A study from Stanford University states that computers determine the sexual orientation of a person using facial detection technology.[27] This study shows that a computer algorithm could differentiate between gay and straight men 81 per cent of the time and 74 per cent for women.[28] Kosinski states that face-analysing algorithms can deduce gay or straight people by their photographs. This facial detection algorithm identified the sexuality of men and women on a dating site with up to 91 per cent accuracy.[29] Gary worries that AI will fail Lesbian, Gay, Bisexual, Transgender, and

21 Caliskan, Bryson and Narayanan 2017.

22 Charlie Pownall, 'Incident Number 6' in S McGregor (ed), *Artificial Intelligence Incident Database*. Partnership on AI. (2016).

23 Ari Schlesinger, Kenton P O'Hare and Alex S Taylor, 'Let's Talk About Race: Identity, Chatbox, and AI' <https://static1.squarespace.com> last accessed 22 August 2021.

24 Katie Paul and Akanksha Rana 'U.S Charges Facebook with Racial Discrimination in Targeted Housing Ads' <https://www.reuters.com/article/us-facebook-advertisers-idUSKCN1R91E8> last accessed 19 September 2021.

25 Adam Gabbat, 'Facebook Charged with Housing Discrimination in Targeted Ads' (2019) <https://www.guardian.com> last accessed 30 August 2021.

26 Sweeney (2013).

27 'Row Over AI that "Identifies Gay Faces"' (2017) <https://www.bbc.com> last accessed 13 June 2021.

28 Alex Sharpe and Senthorun Raj, 'Using AI to Determine Queer Sexuality is Misconceived and Dangerous' (2017) <https://theconversion.com/using-ai-to-determine-queer-sexuality-is-misconceived-and-dangerous-83931> last accessed 22 August 2021.

29 Sam Levin, 'New AI can Guess Whether you're Gay or Straight from a Photograph' <https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-you're-gay-orstraight-from-a-photograph> last accessed 19 September 2021.

Questioning (LGBTQ) because it excludes and pushes to the margin data that do not have a robust example.[30] The Stanford University study observed that owing to the grooming styles, gender-atypical features and expressions associated with gay men and women is interpreted that gay men appear more feminine and vice versa.[31] These have raised the ethical question of why stereotypes that libertarians seek to rectify would amplify AI algorithms. It is worrisome that such tools could be used negatively to harm such persons after profiling them based on their gender identity.

AI Discrimination Based on Race

Racial diversity in technology is low, and there has been more talk than work. Little is being done about implementing measures to achieve more diversity in the AI community. Illustrating gender diversity is that the industry is not investing equally in addressing people of colour's imbalance.

Despite the talks to promote diversity, particularly in a race from famous tech companies, a recently released report reveals that Google[32] and Microsoft's[33] share of technical employees who are Black or Latin rose by less than a percentage point since 2014. The share of Black technical workers at Apple [34] has not changed from 6 per cent. Like numerous social structure facets, the AI industry is predominately dominated by white men between 18 and 35 years.[35] Black workers represent 4 per cent of Facebook and Microsoft's entire workforce and 2.5 per cent of Google's.[36]

How training data is edited, organized, altered, etcetera can also encourage discrimination. Such discrimination can happen when historical biases are not corrected before inserting them into the algorithm. This is crucially important if historical datasets had possibly recorded prejudicial actions or discriminatory beliefs into the decision outcomes.

The algorithm can learn human biases through a feedback loop, even if it was initially trained on unbiased data. This can occur if an AI model is used to support human decision-making processes; a human uses the AI model's output and other non-AI methods to arrive upon a final decision outcome, which is fed back into the training data.[37] For instance, law enforcement activities such as predictive policing often target communities of colour, occasioning disparate arrests of people of colour. These arrest figures are logged into the system, which serves as data points

30 Jamie Wareham, 'Why Artificial Intelligence is Set Up to Fail LGBTQ People' (2021) <https://www.forbes.com> last accessed 22 August 2021.

31 Levin (n 29).

32 Google Diversity Annual Report 2020 <https://kstatic.googleusercontent.com/files/> last accessed 20 August 2021.

33 <https://www.microsoft.com/en-us/diversity/inside-microsoft/default.asp> last accessed 20 August 2021.

34 Different Together—20018 Inclusion in Diversity— https://www.apple.com/diversity/ last accessed 19 August 2021.

35 Marcus Thuiller, 'More Diversity Needed to Fight Human Biases' *The Daily Northwestern* (13 November 2019) <https://dailynorthwestern.com/2019//13/oppinion/thiuillier-more-diversity-needed-n-ai-to fight human biases> last accessed 19 August 2021.

36 Ayanna Howard and Charles Isbell, 'Diversity in AI: The Invisible Men and Women' (2020) <https://sloanreview.mit.edu/article/diversity-in-ai-the-invisible-men-and-women/> last accessed 22 August 2021.

37 Thuiller (n 35).

eventually used to create AI systems.[38] Unless something is done to correct this norm, bias will continue to devise future AI systems. For example, the arrests made during Gorge Floyd's protests to stop racist police brutality will amount to data points for future data sets that will be used to build AI predictive policing algorithms.

Buolamwini, a graduate student at MIT in 2015, shared her experience on how some facial analysis software could not detect her dark-skinned face until she wore a white mask.[39] AI algorithms of big tech companies could not accurately detect the faces of iconic black women like Michelle Obama, Oprah Winfrey and Serena Williams.[40] Surprisingly, such famous women cannot be classified by AI systems, which raises ethical questions on how these algorithms are developed.

Apple's iPhone X failed to distinguish between two distant Asian people.[41] Google photo's mistakenly classified black people as gorillas.[42] Jacky Alcine, a Brooklyn resident, discovered that pictures of him and his friend, both blacks, were labelled 'gorillas' in the Google Photos app.[43] These offensive biases in AI can lead to discriminatory or exclusionary practices.

THE IMPACT OF ETHICAL GUIDELINES AND REGULATIONS ON AI

Several countries, companies, research organizations, initiatives[44] and the international regime have publicized their ethical guidelines and AI recommendations. Notably, Canada was the first country to launch a national AI strategy.[45] The Pan-Canadian AI Strategy has four significant goals: 'to develop global thought leadership on the ethical, policy and legal implications of advances in artificial intelligence'.[46] China announced its ethical principles and governance technology development of AI. Its goal is to align the link between AI development and governance and ensure AI's trust and safety, amongst other things.[47] In April 2019, the European

38 Mariam Vogel, 'Biased AI Perpetuates Racial Injustice' (2020) <https://techcrunch.com/2020/06/24/biased-ai-perpetuates-racial-injustice/> last accessed 20 September 2021.

39 Joy Buolamwini, 'Artificial Intelligence Has a Problem with Gender and Racial Bias. Here's How to Solve it' (2019) <https://time.com/5520558/artificial-intelligence-racial-gender-bias/> last accessed 20 September 2021.

40 Stephanie Organ, 'Is AI Sexist and Racist?' (2021) <https://www.sciencefocus.com/> last accessed 22 August 2021.

41 Sophie Curtis, 'iphone X Racism Now: Aplle's Face IA Fails to Distinguish between Chinese Users' <https://www.mirror.co.uk/tech/apple-accused-racism-after-face-11735152> last accessed 16 July 2021.

42 Pete Pachal, 'Google Photos Identified Two Black People as "Gorilla"' <https://mashable.com/2015/07/01/google-photos-black-people-gorillas> last accessed 16 July 2021.

43 ibid.

44 AI initiatives such as the International Standards Organisation- JTC 1-SC42, Information Technology Industry Council (ITIC); Promoting Responsible Development and Use from the report AI Policy Principles, International Telecommunication Union (ITU), Institute of Electrical and Electronics Engineers (IEEE); Ethically Aligned Design, AI Now Institute, UN Centre for Artificial Intelligence and Robotics and Partnership on AI amongst others.

45 Samir Saran, Nikhila Natarajan and Madhulika Srikumar, 'AI and National Straegies' <https://orfonline.org> last accessed 13 July 2021.

46 'Pan-Canadian Artificial Intelligence Strategy' <http://www.jaist.ac.jp/~bao/AI/OtherAIstrategies/Pan-Canadian%20Artificial%20Intelligence%20Strategy.pdf> last accessed 5 August 2021.

47 Wenjun Wu, Tiejun Huang and Ke Gong 'Ethical Principles and Governance Technology Development of AI in China' (2020) <https://www.sciencedirect.com/science/article/pii/S2095809920300011> last accessed 19 September 2021.

Commission High-Level Expert Group on Artificial Intelligence released the final version of its 'Ethics Guidelines for Trustworthy Artificial Intelligence'.[48] Leading multinational corporations such as Microsoft, Apple, Amazon, Facebook, Google and IBM have also launched principles for future AI developments. These ethical guidelines and frameworks are non-binding, and they have recorded negligible success rates in regulating the development and use of AI. It is essential to discuss some of these ethical and regulatory measures on AI.

Ethical and Regulatory Measures Surrounding AI

European Non-Discrimination Law

The European Non-Discrimination Law is one of the significant laws that can be used to protect against discrimination in the context of AI.[49] The poser now is, can this law respond effectively to the ethical bias surrounding AI?

Article 14, European Convention on Human Rights provides that:

> The enjoyment of the rights and freedoms sets forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.[50]

The European Convention outlaws both direct and indirect discrimination.[51] Direct discrimination relates to discrimination of persons based on protected characteristics like racial origin.[52] Indirect discrimination involves implicit discrimination based on a proscribed ground that results in an unequal effect associated with such proscribed ground.[53] It is usually embedded in rules or policies that apply to everyone but unfairly impacts people who share a distinct attribute.[54]

The key difference between direct and indirect discrimination lies in the façade of neutrality.[55] In cases of direct discrimination, the unequal dealings between two persons are based on the prohibited ground; it is the easily spotted form of discrimination. While, in indirect discrimination, the difference of treatment is not recognisably allied

48 Christian Guttmann, 'An Overview of Artificial Intelligence Ethics and Regulations' (2019) <https://medium.com/@ChrisXtg/an-overview-of-artificial-intelligence-ethics-and-regulations-917859fdcb77> last accessed 6 August 2021. See also, the GDPR, the European Parliament resolution of 2017, Senate Bill No. 1001 of the California government etcetera.

49 The Data Protection Law is another instrument that can be successfully used.

50 Protocol 12 of the Convention expands this prohibition to safeguard against discrimination in any legal right provided in national law, even if such legal right is not covered under the Convention.

51 Case law ECHR.

52 Frederick J Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence' (2020) The International Journal of Human Rights DOI:10.1080/13642987.2020.1743976

53 Oran Doyle, 'Direct Discrimination, Indirect Discrimination and Autonomy' (2007) 27 Oxford Journal of Legal Studies <https://www.jstor.org/stable/4494598> last accessed 6 August 2021.

54 'Indirect Discrimination/Australian Human Rights Commission' <https://humanrights.gov.au/quickguide/12049> last accessed 6 August 2021.

55 'Indirect Discrimination' <https://www.eurobound.europa.eu> last accessed 6 September 2021.

to proscribed grounds.[56] It occurs when policies, practices or rules are the same for everyone but has an unfair effect on those with specific protected characteristics.[57] Indirect discrimination must not involve any discriminatory intent so long as 'disadvantage to the protected person results from any identifiable practice, the absence of a legitimate justification for that practice will be taken to constitute discrimination'.[58] Criminal law cannot punish perpetrators of indirect discrimination because the facts establishing the asserted discrimination does not show clear intent to treat a person less favourably based on prohibited grounds.[59] It is common knowledge that intent is an essential element to prove the commission of a crime.

Although the Non-discrimination Law can help safeguard people against AI-related discriminations, it has some drawbacks. Often, AI-related discrimination falls within indirect discrimination; this shows that it would be difficult to prove the AI developer or user's intention. AI systems are usually black boxes, opaque and complex to understand what they do and how they function.[60] The opacity of AI systems makes it difficult the decipher direct discrimination.

Also, the prohibition of indirect discrimination does not provide a plain rule.[61] The ban does not apply if the suspected discriminator successfully prays an objective justification. The European Court of Human Rights provides that:

> A general policy or measure that has disproportionately prejudicial effects on a particular group may be considered discriminatory even where it is not specifically aimed at that group, and there is no discriminatory intent. This is only the case; however, if such policy or measure has no 'objective and reasonable justification.[62]

The alleged discriminator will escape liability from the foregoing if he/she proves that the policy or measure justification is objective and reasonable.

Data Protection Law

The General Data Protection Regulation (GDPR) and Convention contain various rules that safeguard persons against unfavourable and unlawful discrimination.[63] For instance, Article 22(1) of the GDPR provides that:

56 'Indirect Discrimination: Equality & Diversity' <https://www.equality.admin.cam.ac.uk> last accessed 22 August 2021.

57 'Indirect Discrimination: Australian Human Rights Commission' <https://humanrights.gov.au/> last accessed 22 August 2021.

58 European Anti-Discrimination Law Review: The Atlantic Philanthropies' <https://www.atlanticphilanthropies.org> last accessed 6 September 2021.

59 Kasper Lippert-Rasmussen, 'Punishment and Discrimination' in J Ryberg and JA Corlett (eds) *Punishment and Ethics* (Palgrave Macmillan 2010).

60 Carlos Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence'; J. Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algoritms' (January 2016) Big Data Soc 3 <https://doi.org/10.1177/2053951715622512> last accessed 13 June 2021.

61 Borgesius (n 2).

62 *ECtHR, Biao v Denmark* (Grand Chamber), No 38590/10, 24 May 2016.

63 See Recital 71 GDPR on Profiling.

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

However, Article 22(2) provides exceptions to this rule. It provides thus:

> Paragraph 1 shall not apply if the decision:
>
> a. is necessary for entering into, or performance of, a contract between the data subject and a data controller;
> b. is authorized by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
> c. is based on the data subject's explicit consent.[64]

If the decision is crucial for a contract between the data subject and the controller, it is authorized by law or founded on the individual's consent. The prohibition provided in Article 22(1) does not apply.

Jacob Turner opines that the GDPR is not mainly targeted at AI, but its provisions would likely have far-reaching impacts on the industry beyond the drafter's intentions.[65] Also, Borgesius argues that data protection law is not enveloping enough to safeguard people against algorithm discrimination. He highlights the drawbacks of such laws to include compliance and enforcement and their applicability to personal data.[66] This translates that data protection laws apply to an individual rather than groups. Some other limitations centre on the difficulty to explain the reasoning behind a decision since the decision is grounded on extensive amounts of data.

The Organization for Economic Co-Operation and Development

In May 2019, the Organization for Economic Co-Operation and Development (OECD) launched its five fundamental principles for AI operation, which was adopted by 42 countries. It is the first AI principle signed up by governments including non-OECD members, such as Brazil, Costa Rica, Argentina, Malta, Peru, Romania and Ukraine.[67] It also provided recommendations for practical use by governments. The OECD AI Principles are as follows:

1. Inclusive growth, sustainable development and well-being: Stakeholders are encouraged to proactively engage in stewardship of AI in order to pursue beneficial outcomes.

64 art 22 (1) GDPR, <https://gdpr.eu/article-22-automated-individual-decision-making/> last accessed 22 September.

65 Turner (n 4) 229.

66 Borgesius (n 52).

67 'Artificial Intelligence' OECD Principles on AI, <https://www.oecd.org/going-digital/ai/principles/> last accessed 19 September 2021.

2. Human-centred values and fairness: Organizations and individuals that deploy and develop AI should respect the rule of law, human rights, and democratic values.
3. Transparency and explainability: AI market participants should commit to 'transparency and reasonable disclosure' for AI systems. This involves fostering global awareness and understanding of AI systems and ensuring those affected by an AI system understand the outcome and can challenge the outcome if they disagree.
4. Robustness, security and safety: Stakeholders should focus on risk management across the AI lifecycle to ensure systems function appropriately and do not pose a safety risk. In particular, developers should think about the traceability of data, processes, and decisions so that AI outcomes can be analysed.
5. Accountability: Throughout the AI lifecycle, actors are accountable for the proper functioning of AI systems.[68]

The principles are meant to supplement each other and serve as a foundation for strategy development in the future. It is important to stress that the principle is not legally binding.

Additionally, although various enactments like the Consumer, Competition Law, the Civil Rights Act and the Fair Housing Act can be used to defend people against AI discrimination, these laws are not sufficient. The use, benefit and challenges posed by AI cuts across borders. International legislation is fundamental to govern the creation and use of AI to address AI's ethical bias.

ALTERNATIVE VIEWS ON EFFORTS TO MITIGATE AI RISKS

There are alternative views on mediums by which discriminatory AI risks could be mitigated aside from AI ethical guidelines and recommendations. Researchers and AI ethicists have proposed efforts to reduce AI risks at the design stage. Okyere-Manu advocates that African cultural values should be borne in mind at innovative technology's design and development stages.[69] It is common knowledge that developed countries with western cultural backgrounds design the most innovative technological models, and some aspects of these technologies may compromise the African value system.[70] The point foregrounded is appreciated; however, all marginalized groups' perspectives should be considered when innovative technological systems are designed and created.

Hildebrandt suggests the legal Protection by Design (LPbD) approach, which targets legal protection into the computer or network technology code and data-driven

68 Charles S Morgan, 'OECD Principles on Artificial Intelligence Released' (2019) <https://www.lexology.com/library/detail.aspx?g=2ed8febd-0a43-4494-88fd-a6f1fb2f3ed4> last accessed 6 August 2021.

69 Beatrice Dedaa Okyere-Manu, 'Introduction: Charting an African Perspective of Technological Innovation' in Beatrice Dedaa Okyere-Manu (ed), *African Values, Ethics, and Technology: Questions, Issues, and Approaches* (London: Palgrave Macmillan 2021).

70 Malesela John Lamola, 'Africa in the Fourth Industrial Revolution: A Status Quaestionis, From the Cultural to the Phenomenological' in Okyere-Manu, ibid.

ecosystem.[71] The articulation of legal protection is translated into code at the design stage. Such formalization ensures the logical operation of deduction that is vital for automation. LPbD is somewhat similar but not identical with 'safer by design'.[72] The goal of LPbD is to 'seek ways to prevent diminished legal protection by intervening in the design of the relevant computational architecture, where design refers to the joint constructive work of whoever make, build, assembly, and construct such architectures'.[73] LPbD is a good technological measure that can somewhat mitigate AI risk. The question is, who will ensure that such legal protection is embedded at the design stage? The approach is unclear on who should take action if such protection is not translated at the design stage. These questions bring us to the need for AI vigilantes who carry out the model evaluation at the design stage and contribute to regulating AI to suit their particular circumstances.

A CASE FOR AILANTISM AND GLOBAL AI GOVERNANCE

It is laudable that national, international efforts, industry-based ethical guidelines, principles and recommendations have been launched. The majority of these ethical guidelines cover fairness, human-centred values, explainability, accountability, etcetera. However, its non-binding effect creates a soft landing for violators. The enforcement of AI ethical standards 'may involve reputational losses in the case of misconduct, or restrictions on memberships in certain professional bodies'.[74] Internal self-governance is not sufficient; a legally binding global framework should be implemented to ensure the proper development and use of AI. Also, a unifying global framework is critical to regulating AI across the board to avoid chaos with multiple industry self-regulation.

It is vital to employ diverse representation in AI's design, development, deployment and governance to avoid abuse and technological harm scenarios. The low testing or selection of marginalized groups in data that shapes AI has resulted in technological inventions based on a small-scale fragment of the world. These inventions do not represent a thorough analysis of different groups across the globe. Marginalized communities should be engaged in the development and governance of AI.

All AI tools and models are created by human beings. Logically, machine learning models reflect the biases of developers, organizational teams, data scientists who implement the models and the data engineers that gather data.[75] Human influence cannot be disregarded from data. Datasets are used to train an algorithm. During such training, human activities can introduce bias at the decision stage and how the data is

71 Mireille Hildebrandt, 'Code-Driven Law: Freezing the Future and Scaling the Past' in C Markou and S Deakin (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing 2020) 67–84.

72 Mireille Hildebrandt, 'Saved by Design? The Case of Legal Protection by Design' (2017) 11 Nanoethics https://doi.org/10.1007/s11569-017-0299-0 last accessed 20 September 2021.

73 Hildebrandt (n 71).

74 Thiol Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines' (2020) 30 Minds & Machines 99–120 .

75 Michael Mckenna, 'Machine and Trust: How to Mitigate AI Bias' <https://www.toptal.com/artificial-intelligence/mitigating-ai-bias/> last accessed 18 August 2021.

collected or categorized.[76] Coeckelberg reiterates that machine learning can be supervised. This means 'that the algorithm focuses on a particular variable that is designated as the target for prediction'.[77] Supervised learning is when a supervisor or teacher trains the machine using well-labelled data.[78] If the programmer teaches the model by providing examples and non-examples, the model digests these examples/labels and can predict or categorize new sets of data in future. Algorithms are unsupervised when no examples or labels are given.[79] The model finds its structure in its inputs.[80]

Since human influence is paramount in all stages of AI development, humans will also be influential in mitigating AI risks. AI vigilantism is a concept whereby non-state actors form a gatekeeping network to enforce ethical and legal standards by ensuring AI's fair creation, development, and use. Vigilantism is nothing new; it has been deployed in the past by citizens to safe guard their lives, properties or fight a cause. It is defined 'as the extralegal prevention, investigation, or punishment of offences'.[81] It is referred to as a type of automization and a form of community involvement.[82] Automization' involves a context in which an ideal-typical state claims to monopolize law-enforcement functions, in contrast to groups acting strictly autonomously, or as challengers of state law-enforcement institutions'.[83]

Scholars have condemned vigilantism as undermining the rule of law due to the absence of checks and balances.[84] Kosseff opines that cyber vigilantism can breed undesirable consequences such as abuse of power and uneven punishments.[85] Despite the criticism of vigilantism, some vigilante groups have been praised for assisting law enforcement agents in combating crimes. For instance, online vigilante groups employ paedophile hunting which helps the police uncover real paedophiles.[86]

76 Genevieve Smith and Ishita Rustagi, 'Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook' (2020) <https://haas.berkeley.edu> last accessed 18 September 2021.

77 Coeckelberg (n 1).

78 Shubham Bansal, 'Supervised and Unsupervised Learning' (2021) <https://www.geeksforgeeks.org/supervised-unsupervised-learning/> last accessed 22 September 2021.

79 Sanatan Mishra, 'Unsupervised Learning and Data Clustering' (2017) <https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eecb78b422a> last accessed 19 September 2021.

80 Unsupervised learning has its disadvantages/issues such as its difficulty compared to supervised machine learning and the uncertainty of its results since no examples/answer labels were provided. However, it is needed in cases of large datasets which is costly to annotate and few examples can be labelled manually.

81 Regina Bateson, 'The Politics of Vigilantism' (2020) <https://doi.org/10.1177/0010414020957692> last accessed 13 September 2021.

82 Gilles Favarel-Garrigues, Samuel Tanner and Daniel Trottier, 'Introducing Digital Vigilantism' (2020) <https://www.tandfonline.com/doi/full/10.1080/17440572.2020.1750789> last accessed 13 September 2021.

83 ibid.

84 Joseph Steinberg, 'We Must Condemn, Not Celebrate, The Download of Parler's Data: Hacker Vigilantism May Even Help Criminals More Than Law Enforcement' <https://josephsteinberg.com/we-must-condemn-not-not-celebrate-the-download-of-parlers-data-hacker-vigilantism-may-even-help-criminals-more-than-law-enforcement/> last accessed 13 August 2021.

85 Jeff Kosseff, 'The Hazards of Cyber-Vigilantism' (August 2016 Computer Law & Security Review vol 32, issue 4 <https://doi.org/10.1016/j.clsr.2016.05.008> last accessed 13 September 2021. Koseff added that vigilante groups should collaborate with governments and other private actors to combat cyber threats rather than working independently.

86 Mark Button, 'Vigilantes and Private Security are Policing the Internet Where Government have Filed' (2020) <https://theconversation.com/vigilantes-and-private-security-are-policing-the-internet-where-government-have-failed-132040> last accessed 13 September 2021.

The emergence of Allantism would come in handy to mitigate harms, risks and abuse associated with AI by creating laws regulating the designing and use of AI. The AI vigilantes will police the use of AI by humans and report any abuse or harm suffered to law enforcement agents. AI vigilante's role will be to monitor AI's creation and use it to ensure that the developers are accountable for any harm suffered by humans. Although vigilantism has some disadvantages that may lead to negative consequences due to policing methods, [87] it positively impacts if implemented in cooperation with law enforcement agencies.

They will work closely with law enforcement agencies and the government to avoid negative consequences like undermining democratic systems' lawfulness and alleged abuse of the system. In South Africa, a vigilante group from the Port Elizabeth's township's was incorporated as an official 'Safety and Security' structure under the Community Policing Forum'.[88]

The appointed AI vigilantes will comprise persons/groups with an increased risk of their rights being disproportionately impacted by AI. In recent times, there have been increased activism of indigenous communities developing guidelines for indigenous knowledge. Many institutions have ethical guidelines and research protocol agreements. Also, many indigenous organizations have established principles and protocols that researchers are required to follow. For instance, the First Nations communities have ownership over their information and cultural knowledge; they equally have control over how their data is used or accessed and must be consulted and give consent to all stages of the research style.[89]

Jose R Martinez Cobo's Study on the Problem of Discrimination against Indigenous Populations provided one of the most cited descriptions of the concept of 'indigenous' thus:

> Indigenous communities, peoples and nations are those which, having a historical continuity with pre-invasion and pre-colonial societies that developed on their territories, consider themselves distinct from other sectors of the societies now prevailing on those territories or parts of them. They form at present non-dominant sectors of society and are determined to preserve, develop and transmit to future generations their ancestral territories, and their ethnic identity, as the basis of their continued existence as peoples, in accordance with their own cultural patterns, social institutions and legal system.[90]

The United Nations recognized that indigenous people are 'arguably among the most disadvantaged and vulnerable groups of people in the world'.[91] A formal/

87 For instance, the use of methods like organized denunciation, shaming, hounding, flagging and doxing in digital vigilantism has been greatly criticized.

88 Lars Buur, 'Democracy & Its Discontents: Vigilantisms, Sovereeignty & Human Rights in South Africa', Review of African Political Economy, 35 DOI:10.1080/03056240802569250 last accessed 22 September 2021.

89 The First Nations Information Governance Centre help achieve indigenous data sovereignty.

90 'The State of the World's Indigenous Peoples' <https://www.un.org/esa/socdev/unpfii/documents/SOWIP/en/SOWIP_web.pdf> last accessed 26 July 2021.

91 'Indigenous Peoples at the United Nations' <https://www.un.org/development/desa/indigenouspeoples/> accessed 22 August 2021.

universal definition of indigenous people at the international level to be adopted by states was declined by observers from indigenous organizations and government delegations.[92] Instead, Article 33 of the United Nations Declaration on the Rights of Indigenous Peoples highlights the significance of self-identification. It provides as follows:

Article 33

1. Indigenous peoples have the right to determine their own identity or membership in accordance with their customs and traditions. This does not impair the right of indigenous individuals to obtain citizenship of the State in which they live.
2. Indigenous peoples have the right to determine the structures and to select the membership of their institutions in accordance with their own procedures.[93]

From the preceding provision, indigenous people have the right and should play an active role in determining their institutions' structure and membership.

Article 2 of the UN Declaration on the Rights of Indigenous People states that 'indigenous peoples are free and equal to all other peoples and individuals and have the right to be free from any kind of discrimination, in the exercise of their rights, in particular, that based on their indigenous origin or identity'.[94] Article 15 (2) provides as follows:

> States shall take effective measures, in consultation and cooperation with indigenous peoples concerned, to combat prejudice and eliminate discrimination and to promote tolerance, understanding and good relations among indigenous peoples and all other segments of society.

Article 15 of the UN Declaration on the Rights of Indigenous People upholds indigenous people's full and active involvement in matters that concern and negatively affect them. The full participation of marginalized groups in shaping a universal AI regulation will mitigate the global ethical issue surrounding AI.

CONCLUSION

Minority and marginalized groups have suffered grave discriminatory and technological risks due to AI ethical dilemmas. This study examined the ethical measures and guidelines on AI. It found that the non-binding effect of such standards and policies acts as a soft landing for violators since they set no imminent threat or punishment. AI creators should be responsible for the risks and harms perpetuated by its creation. It is essential to have a legally binding and universal framework that regulates AI technology. This article explores the deployment of AIlantism, an initiative distinct from the usual black letter law, to aid in developing a global AI framework.

92 'The State of the World's Indigenous Peoples'

93 art 1 of the ILO Convention No 169 also underlines the weight of self-identification.

94 art 2 UN Declaration on the Rights of Indigenous People.

AI vigilantes will monitor AI technology development, design, programming and operation to reduce future risks. They will work closely with the government and the police to ensure checks and balances. Such involvement in the creation of various AI technological systems will reduce human biases learnt by algorithms.

CONFLICTS OF INTEREST SECTION

The author of the article, "AI Ethical Bias: A Case for AI Vigilantism (AIlantism) in Shaping AI Regulation", declares that there is no conflict of interest.

This work has no funding or support from any organisation or institution.