Artificially Intelligent and Inclusive by Design: A Human-Centered Approach to Online Safety

Daricia Wilkinson

Clemson University, 821 McMillan Rd, Clemson, SC 29631, United States

Abstract

Social media platforms have played a key role in the amplification of online risks and harms. Moreover, the disproportionate rates of victimization for gender and sexual minorities (GSM) have raised growing concerns since the consequences of online harms extend well beyond the platforms on which they occur. This is particularly complex in non-Western contexts, like the Caribbean, where varying views on norms, culture, and legislative protection further highlight the need for the careful design of approaches that might be automated. This position paper discusses practical and ethical questions related to the inclusive design of AI-supported safety mechanisms in social media by proposing systems that are artificially intelligent and inclusive by design.

Keywords

social media, artificial intelligence, fairness, inclusion, design, caribbean

1. Introduction and Background

In 2020, there were 4.2 billion active social media users, and on a daily average worldwide, a person would use social media for about 145 minutes [1]. Undoubtedly, social media has persistently evolved to become an ever-present force that allows billions of people to connect, communicate, and learn from others. In the same light, the explosive rate of adoption has been accompanied by failures to reduce the relentless harms that continue to persist on these platforms.

While these platforms have been vital in maintaining safety, they have also served as highways for Mis, Dis, and Mal-information. Evidence of this dichotomy surfaced during the La Soufrière eruption in St. Vincent where many citizens used Facebook's safety tool to inform loved ones that they were not harmed while the same platform was used to a tool to spread misinformation throughout the region about other volcanoes potentially erupting [2].

The scale and variety of harms have motivated social media companies to harvest the power of artificial intelligence (AI) approaches to prevent, detect, and rectify harms. Some advocates have propped AI as a panacea that would identify misinformation, hate speech, pornographic material, and other platform violations in a quick and fair manner before the offending content is uploaded for others to see. This idea was unofficially tested on a large-scale during the coronavirus pandemic as many social media companies relied on automated content moderation as their human moderators were sent home. It was a failure. Human rights journalists who rely on

AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA

dariciw@clemson.edu (D. Wilkinson)

social media to document injustices, saw multiple accounts of activists being shut down without the option to appeal the decision [3]. Meanwhile, problematic content remained untouched as human moderators were not able to serve as arbitrators and determine the nuances that would indicate platform violation. Consequently, the numbers for the removal of high risk content, like child exploitation and self-harm on Facebook, were 40% lower in the second quarter of 2020 [3]. The results of this test raise questions about the reliance on solely automated approaches. Is it fair to rely on technology to decode complex human issues that humans have difficultly with - especially when the matters include systematic oppression, race relations, political power-plays, and economic dynamics, etc.? In addition, within this domain, the research, policies, and design of safety mechanisms, have been largely dominated by researchers from western, educated, industrialized, rich and democratic (WEIRD) nations [4] with imbalances regarding the data sets used in models and systems as well as the representation of socio-cultural groups.

This paper explores the opportunities for an alternate perspective that departs from the one-size-fits-all approach to safety mechanisms for social media - systems that are *artificially intelligent and inclusive by design* (AIIbD). For this paper, I consider the Caribbean as a use-case since the region is rich in cultural diversity while also having unique challenges that would require a thoughtful approach for AI transformation. Arguably, applying this approach to safety mechanisms creates its own set of ethical questions. In the subsequent sections, this paper provides an overview of the use of AI-support safety mechanisms. Next, I outline ethical questions regarding inclusive AI-support safety mechanisms. Finally, I conclude by discussing challenges and opportunities for future research.

2. AI-Supported Safety Mechanisms

Various machine learning techniques have been developed and implemented to help reduce the risks to online safety in social media [5] [6] [7] [8] [9]. Platforms have implemented both proactive and reactive approaches. Reactive systems are triggered when a user already identifies problematic content which is brought to the platform's attention and it is then evaluated based on the policies and standards of that platform. The effectiveness of these methods have been criticised since the success of the technique is reliant on users flagging content. This may increase the possibility of risky content being circulated before finally being flagged. With reactive approaches, options for justice could include punitive approaches such as removing content or banning users [10].

On the other hand, proactive approaches can include both manual and automated approaches. These may include, delaying the publication content until they are evaluated by a human, the use of filters that prevent potentially problematic content from being posted, evaluating posting behavior to proactively block spam, or network-level signals such as IP addresses. Proactive techniques have been used, for example, in the detection of potentially illegal objects in images and to reduce the intentional distribution of unsolicited images [11]. Moreover, AI techniques have been deployed to reduce other malicious activity such as the prevalence of fake news (see [12] for an overview). AI-supported safety mechanisms have similar characteristics in that their successful deployment is dependent on big but diverse data sets. Thus, how can we apply such techniques to the new types of harms that will be culturally-responsive?

2.1. Beyond a one-size-fits-all approach

As detailed above, there are several techniques and approaches that could assist in the technical development of AI-supported safety mechanisms. The rate of progress in the field of artificial intelligence is exponentially increasing and what may seem like a distant reality could quickly evolve into mature technology. Although the power of AI is acknowledged, it is also equally important to consider the people for which it is being designed. Within system design, diversity is an important principle that helps to illuminate problems from invisible communities and create powerful solutions. Thus, this paper considers how we could combine the power of AI and the power of diversity, and proactively develop principles for artificially intelligent and inclusive safety mechanisms.

3. Author's Positionality Statement

This paper explores Caribbean cultural contexts and culturally responsive approaches for Human-Computer Interaction (HCI). As such, the work of minoritized scholars has informed the perspectives and approach taken in this study [13]. The researcher conducting this study was guided by calls for researchers to be transparent in their positionalities, personal histories, and perspectives in order to conduct collaborative culturally responsive research that will benefit Caribbean communities.

The author is a Caribbean native who resides in the United States to pursue education. Her motivation for research is grounded in a desire to improve the user experience and well-being for social media users in ways that sustain and advance marginalized communities. She collaborates with a non-profit organization in the Eastern Caribbean to encourage the participation of diverse voices.

4. Artificially Intelligent and Inclusive by Design

In the following section, I discuss ethical considerations framed by key questions that would affect the implementation of inclusive AI-supported safety mechanisms within the social media ecosystem using the Caribbean as a use-case.

4.0.1. How do we incorporate inclusivity?

Creating solutions that are artificially intelligent and inclusive by design would inherently require a departure from personal biases and embrace principles such fairness, transparency, and equity. However, there are still improvements to be made for those responsible for the design and development of social media platforms. Myers recently highlighted the imbalance at major social media companies including Facebook and Google [14]. The study revealed that women make up only 15% of AI researchers at Facebook and just 10% at Google [14]. Hence, to avoid mirroring that problem, stakeholders would need to consider protected and marginalized groups, as well as longstanding contexts that might be overlooked. This is particularly important when designing to combat harms. Contextual knowledge would inform design that would help users from different

backgrounds facing the same threat. AIIbD solutions would be knowledgeable of users' context to suggest the best course of action that could assist in seeking a resolution.

Scholars have acknowledged that various cultural, religious, social, and philosophical factors shape boundaries that contribute to what it means to be private which could inherently influence privacy and safety practices in the online space [15, 16, 17]. Prior work show that disclosure on social media platforms follow a *privacy calculus* where users interpret the risks and benefits associated with their activities on respective platforms [18]. Cultural norms have been shown to be a significant predictor of online disclosure [16, 19, 20]. Persons from cultures that are individualist - like the US and Australia - tend to be more concerned about their personal data being misused and under surveillance whereas cultures that are collectivist - like the Caribbean - tend to be more concerned about the harm they could impose on their collective's (e.g. friends and family) privacy [21]. Thus, perceptions and attitudes towards what it means to be safe may vary greatly compared to western contexts. Therefore, the scope of AIIbD safety mechanisms could be expanded to acknowledge that the pursuit of safety within the online space might be entangled with harms traditionally labelled as cyber-crimes (e.g. online harassment or online fraud), physical threats (e.g. both online and offline stalking as a result of online interaction), and social implications.

Therefore, questioning how AI could help Caribbean citizens to address online threats would overlap with understanding historical challenges that might have first existed in the physical world. For example, a response to harassment in the region that is AIIbD would require a holistic investigation of perceptions of offline harassment to understand how AI could assist in reducing the risks associated with this threat. A researcher might consider non-technical factors such as culture as persons from collectivist cultures like the Caribbean may consider direct conflict resolution before escalating to legislative options. Thus, AI-supported solutions could be built to be inclusive of personal differences.

4.0.2. How should systems be policed?

Throughout the Caribbean region, there are ongoing efforts by various stakeholders to develop frameworks that would govern the way AI is used and regulated [1]. However, governance regarding online data protection in the Caribbean are at varying stages of development with a noteworthy number of countries throughout the region without relevant laws in place. Out of the 26 countries in the region, there are only 10 countries with enforced legislation that could provide legal protections in the event of digital harms (see the Table 1). Moreover, many countries within the region enforce laws that require revision for consistency with updated online data protection and safety principles (see [22] for a regional legislative comparison).

As such, even if AI-supported tools are available for citizens to use, in the event of any harm there might be varying options available for justice for persons who live only miles apart.

Additionally, it is important to be culturally sensitive in the design of AI-supported resolutions as victims might not want to highlight their abusers for fear of further victimization and repercussions to their family reputation. To address these issues, AIIbD developers and governing bodies could consider alternative approaches for justice that may include fines, public apologies,

[1] UNESCO led effort: https://ai4caribbean.com

Country	Law	Passed
Antigua and Barbuda	Data Protection Act, 2013 No 10 of 2013	2013
Bahamas, The	Data Protection (Privacy of Personal Information) Act, CH.324A	2003
Barbados	Barbados Data Protection Act, 2019-29	2019
Bermuda	Personal Information Protection Act, 2016 : 43	2016
Cayman Islands	The Data Protection Law, 2017 (LAW 33 OF 2017)	2017
Jamaica	Data Protection Act, 2020 (No 7-2020)	2020
St Kitts and Nevis	Data Protection Act, 5 of 2018	2018
Saint Lucia	Data Protection Act, No 11 of 2011	2011
St Vincent and The Grenadines	Privacy Act of 2003	2003
Trinidad and Tobago	Data Protection Act, 2011	2011

Table 1
The jurisdictions with substantive laws to govern the protection of personal data. 10 jurisdictions have passed data privacy laws.

or community service [10].

4.0.3. What safeguards should be implemented?

Democracies are often led by persons that people trust will make decisions that are in the best interests of the citizens they serve. However, there are instances where that trust is abused. For instance, let us assume that a framework is created that details how violations to AI-supported safety mechanisms are enforced. Safeguards should be implemented to prevent any particular governing or political body to unfairly use the framework to target the opposition to stifle freedom of speech. Likewise, regulations should not be used to to perpetuate further harm. Saki and Sambuli describe how, in Uganda and Brazil, respectively, anti-pornography laws and defamation lawsuits have been used to punish women for being online rather than protecting them [23]. These are critical considerations for keeping gender and sexual minorities (GSM) in the region safe.

Although the points raised are not exhaustive, they might serve as a starting point to help contextualize the needs of people from different backgrounds and highlight ways that AI can support safety and well-being in social media.

4.0.4. Acknowledgements

This work is supported by a FRIDA grant titled *"Online Safety Tools for Vulnerable Groups in the Caribbean"*. I would like to thank the members of HOPE Nevis, a non-profit organization in the Caribbean, which leads the efforts of this grant and Dr. Bart Knijnenburg for his input and discussions.

References

[1] Statista, Social media usage worldwide, 2021. URL: https://www.statista.com/topics/1164/social-networks/.

[2] R. Andrews, St. Vincent's Volcanic Eruption: Misconceptions Debunked And Questions Answered, 2021. URL: https://www.forbes.com/sites/robinandrews/2021/04/12/st-vincents-volcanic-eruption-misconceptions-debunked-and-questions-answered/, section: Science.

[3] M. Scott, L. Kayali, What happened when humans stopped managing social media content, 2020. URL: https://www.politico.eu/article/facebook-content-moderation-automation/.

[4] J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world?, Behavioral and brain sciences 33 (2010) 61–83.

[5] E. Raisi, Weakly Supervised Machine Learning for Cyberbullying Detection, Ph.D. thesis, Virginia Tech, 2019.

[6] J. Li, Q. Xu, N. Shah, T. K. Mackey, A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study, Journal of medical Internet research 21 (2019) e13803.

[7] Y. Kou, X. Gui, Mediating community-ai interaction through situated explanation: The case of ai-led moderation, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–27.

[8] E. J. Llansó, No amount of "ai" in content moderation will solve filtering's prior-restraint problem, Big Data & Society 7 (2020) 2053951720920686.

[9] T. Gillespie, Content moderation, ai, and the question of scale, Big Data & Society 7 (2020) 2053951720943234.

[10] S. Schoenebeck, O. L. Haimson, L. Nakamura, Drawing from justice theories to support targets of online harassment, new media & society (2020) 1461444820913122.

[11] R. M. Hayes, M. Dragiewicz, Unsolicited dick pics: Erotica, exhibitionism or entitlement?, in: Women's Studies International Forum, volume 71, Elsevier, 2018, pp. 114–120.

[12] A. K. Cybenko, G. Cybenko, Ai and fake news, IEEE Intelligent Systems 33 (2018) 1–5.

[13] D. Thakur, How do icts mediate gender-based violence in jamaica?, Gender & Development 26 (2018) 267–282.

[14] S. Myers West, Discriminating systems: Gender, race and power in artificial intelligence (2020).

[15] M. G. Hoy, G. Milne, Gender differences in privacy-related measures for young adult facebook users, Journal of interactive advertising 10 (2010) 28–45.

[16] Y. Li, A. Kobsa, B. P. Knijnenburg, M.-H. C. Nguyen, et al., Cross-cultural privacy prediction., Proc. Priv. Enhancing Technol. 2017 (2017) 113–132.

[17] M. Madden, Privacy management on social media sites, Pew Internet Report 24 (2012) 1–20.
[18] T. Dienlin, M. J. Metzger, An extended privacy calculus model for snss: Analyzing self-disclosure and self-withdrawal in a representative us sample, Journal of Computer-Mediated Communication 21 (2016) 368–383.
[19] H. Krasnova, N. F. Veltri, Privacy calculus on social networking sites: Explorative evidence from germany and usa, in: 2010 43rd Hawaii international conference on system sciences, IEEE, 2010, pp. 1–10.
[20] H. Krasnova, N. F. Veltri, O. Günther, Self-disclosure and privacy calculus on social networking sites: the role of culture, Business & Information Systems Engineering 4 (2012) 127–135.
[21] S. Trepte, L. Reinecke, N. B. Ellison, O. Quiring, M. Z. Yao, M. Ziegele, A cross-cultural perspective on the privacy calculus, Social Media+ Society 3 (2017) 2056305116688035.
[22] A. Bleeker, Creating an enabling environment for e-government and the protection of privacy rights in the caribbean: A review of data protection legislation for alignment with the general data protection regulation (2020).
[23] V. Sika, N. Sambuli, Ict4governance in east africa, in: International Conference on e-Infrastructure and e-Services for Developing Countries, Springer, 2014, pp. 175–179.