Control and Resistance: The Dynamic Interplay of Censorship and Circumvention

RUYUAN WAN and KARLA BADILLO-URQUIOLA, University of Notre Dame, USA

The increasing use of biased algorithmic censorship has led to people's circumvention behaviors and resistance to unfair control. This has made it harder to achieve a balance between user safety and free expression in online spaces. In this paper, we identify current limitations of online censorship and explore the dynamic bi-directional interaction between censorship and user behavior using a case analysis of youth online safety.

1 INTRODUCTION

Censorship has become a predominant approach to combat harmful online behavior such as hate speech and harassment. However, biased algorithms and poorly designed policies that regulate censorship often lead to the ban of non-problematic content or the failure to detect harmful content. This, in turn, can trigger circumvention behaviors by users seeking to evade censorship control, which can further complicate the balance between free expression and user safety. To address these challenges, we plan to investigate the complex and dynamic interplay between censorship and user behavior to develop value-sensitive solutions. We hope to engage with the CHI community and workshop to contribute insights into the social and technical implications of combating online toxicity.

2 ALGORITHMIC CENSORSHIP

To combat toxicity in online social spaces, censorship is an essential approach that typically involves the use of detection algorithms [5]. However, these algorithms have limitations due to inherent biases, leading to false positives and false negatives that adversely affect user trust and platform engagement. Detection algorithms are usually rule-based or machine learning-based. Rule-based algorithms use predetermined criteria, such as keywords, to identify harmful content, but may miss more nuanced forms of toxicity [5]. On the other hand, machine learning-based algorithms use data-driven models to identify harmful content by learning patterns associated with toxicity. Although machine learning-based algorithms can identify harmful content more accurately, they also may not fully capture the complexity of human language and behavior [3, 11]. Moreover, the effectiveness of algorithms is limited because binary confusion matrices [1] may not be able to effectively account for the nuanced nature of toxic content. What may be considered offensive to one group may not be offensive to another. Recent research has explored the use of disagreement and fairness confusion matrices to improve the accuracy and fairness of these algorithms [7, 8, 13]. Yet, finding the right balance between false positives and false negatives remains a critical challenge in censorship technology.

3 UNDERSTANDING CIRCUMVENTION BEHAVIOR UNDER CENSORSHIP

Algorithmic censorship can trigger circumvention behaviors that hinder its effectiveness. Common circumvention behaviors include the use of coded language [9, 15], creation of alternate accounts [4], the use of encrypted channels [6, 12], and the adoption of technical tools such as VPNs [10]. The bi-directional interaction between censorship and user behavior can occur at multiple levels: individual, community, and society. At the individual level, users may feel the need to express themselves in ways that goes against the norms and rules of the platform or society. This can be due to a desire for self-expression, or a reaction to perceived injustices. As a result, users may engage in circumvention behaviors

[1] a performance measurement for a machine learning classification that only considers positive or negative results

to bypass censorship and express themselves freely. At the community level, subgroups may develop their own norms and values that diverge from those of the larger stream. These groups may feel oppressed by censorship efforts, and may engage in circumvention behaviors as a way to resist the status quo. At the societal level, censorship can have broader impacts on human rights. Governments and other controlling entities may use censorship as a means of suppressing dissent, or maintaining political power, which can have negative consequences for democracy. Understanding this complex interaction between censorship and circumvention behavior is critical in designing effective and ethical content moderation policies and technologies that can promote free expression while ensuring user safety.

4 VALUE-SENSITIVE SOLUTIONS FOR PROMOTING ONLINE SAFETY FOR YOUTH

To address the limitations of algorithmic censorship, it is important to develop more nuanced approaches to censorship that take into account the unique needs and experiences of people and the complex nature of online social spaces. Youth online safety is a specific case where value-sensitive solutions are essential. For instance, a false negative case regarding youth might occur when harmful content is not flagged and removed by the algorithm because it is not harmful for adults. On the other hand, a false positive case could be when an adult account is mistakenly identified as an underage account leading to unnecessary account suspension. This is often seen in algorithms designed to detect underage users in gaming platforms to avoid game addiction. The algorithm may flag the account due to certain behavioral patterns. In reality, the user may be an adult who has been mistakenly identified as underage, leading to unfair treatment and potentially harmful outcomes. This flawed algorithm also perpetuates negative stereotypes and bias against younger users. That is, the algorithm may have been trained on biased data that assumes younger players are more likely to be bad teammates, leading to age-based discrimination. However, it is also possible that parents give children their adults accounts to help children circumvent the algorithm because they do not have time to monitor them. To address these challenges, we must include youth in the design and development of content moderation policies and technologies, as this can ensure their unique needs and experiences are taken into account [1, 2, 14]. By adopting human-centric algorithm approaches to content moderation, we consider the different viewpoints of all stakeholders [7, 8, 13] and develop more effective and ethical approaches to promoting safe and inclusive online spaces.

5 CONCLUSION

The dynamic interaction between censorship and user behavior calls for a nuanced approach to the design and use of censorship technology. Detection algorithms is critical but must be balanced with user safety and free expression given the needs of vulnerable populations, such as youth online safety. We are looking forward to fostering a community at the workshop developing ethical and effective approaches to promote online safety for all.

REFERENCES

[1] Karla Badillo-Urquiola, Chhaya Chouhan, Stevie Chancellor, Munmun De Choudhary, and Pamela Wisniewski. 2020. Beyond parental control: designing adolescent online safety apps using value sensitive design. *Journal of adolescent research* 35, 1 (2020), 147–175.

[2] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting risky research with teens: co-designing for the ethical treatment and protection of adolescents. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–46.

[3] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 116–128.

[4] Farhan Asif Chowdhury, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha, and Abdullah Mueen. 2021. Examining factors associated with twitter account suspension following the 2020 us presidential election. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 607–612.

[5] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting gender stereotypes: lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[6] Cori Faklaris and Sara Anne Hook. 2017. Attitudes About'Fair Use'and Content Sharing in Social Media Applications. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 171–174.

[7] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[8] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[9] Heng Ji and Kevin Knight. 2018. Creative language encoding under censorship. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*. 23–33.

[10] Moses Namara, Daricia Wilkinson, Kelly Caine, and Bart P Knijnenburg. 2020. Emotional and practical considerations towards the adoption and abandonment of vpns as a privacy-enhancing technology. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 83–102.

[11] Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 991–1000.

[12] Manya Sleeper, William Melicher, Hana Habib, Lujo Bauer, Lorrie Faith Cranor, and Michelle L Mazurek. 2016. Sharing personal content online: Exploring channel choice and multi-channel behaviors. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 101–112.

[13] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. *arXiv preprint arXiv:2301.05036* (2023).

[14] Pamela Wisniewski, Karla Badillo-Urquiola, Mel Stanfill, and Anastasia Salter. 2017. Using Participatory Design to Give Foster Teens a Voice in Designs for Their Own Online Safety. In *Wisniewski, P., Badillo-Urquiola, K.*, Stanfill, M., and Salter, A.(2017)"Using Participatory Design to Give Foster Teens a Voice in Designs for Their Own Online Safety," Extended Abstract presented at the Workshop on Design Methods for Underserved Communities at the 2017 ACM Conference on Comput*.

[15] Eddie Yang and Margaret E Roberts. 2021. Censorship of online encyclopedias: Implications for NLP models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 537–548.