Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

LEVERAGING AI-BASED INTERVENTIONS TO ADDRESS CYBERBULLYING AMONG CHILDREN: A RIGHTS-BASED PERSPECTIVE

Tijana Milosevic
DCU Anti-bullying Centre and ADAPT SFI

Kanishk Verma
ADAPT Centre, DCU Anti Bullying Centre

Samantha Vigil
University of California, Davis

Michael Carter
Boston Children's Hospital, Digital Wellness Lab

Elisabeth Staksrud
University of Oslo

Derek A. Laffan
DCU Anti-bullying Centre

Brian Davis
ADAPT Centre, Dublin City University

James O'Higgins Norman
DCU Anti-bullying Centre

Suggested citation: Milosevic, T., Verma, K., Vigil, S., Carter, M., Staksrud, E., Laffan, D., Davis, B., O'Higgins Norman, J.(2022, November). Leveraging AI-Based Interventions to Address Cyberbullying among Children: A Rights-Based Approach. Paper presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

Introduction

Cyberbullying, broadly defined as repeated harm inflicted with digital technology, continues to pose a significant problem for children. Online intermediaries, such as social media platforms[1], are leveraging algorithmic techniques designed to automate the process of moderation such as natural language processing (NLP), machine and deep learning (artificial intelligence or AI) in order to optimise cyberbullying moderation (Gorwa et al., 2020; Vidgen & Derczynski, 2020). This allows human moderators to process cases faster when they are reported by users; and to detect cyberbullying proactively—before these incidents are reported to the platform and subsequently flagged for moderation.[2]

While one in three Internet users are children, children are still insufficiently consulted or taken into consideration in Internet governance decisions, as well as decisions that relate to platform design (Livingstone & Third, 2017). For example, social media companies do not disclose whether and how they consult children in their safety design decisions. In as much as safety design decisions by social media platforms have implications for children, children need to be consulted in this process, if interventions are to be effective from their perspective; and if they are to balance children's rights to protection vs. participation and privacy. Following the United Nations Convention on the Rights of the Child (UNCRC)[3], which applies in digital environments[4], not only do children have the right to protection (right to safety, such as protection from cyberbullying) but they also have the right to be consulted on matters that concern them and to privacy, among others. Children's rights to protection are often prioritised over their rights to participation and privacy (e.g., they might be denied access to social media features for the sake of safety; or they might be surveilled online by their parents to ensure protection, see Livingstone & Third, 2017; Mascheroni & Siibak, 2021; Staksrud, 2016).

We therefore ask:

RQ1: How can we design automatic tools that support effective proactive bullying interventions that assist victimised children while ensuring children's rights to privacy, freedom of expression and other relevant rights as outlined in the UNCRC?

RQ2: How can we leverage children's feedback to optimise the effectiveness of such tools?

The current study

[1] We are aware of the problematic nature of the term "platform" but we use it here for the lack of a more suitable term see Gillespie, 2017.
[2] https://transparency.fb.com/
[3] https://www.unicef.org.uk/what-we-do/un-convention-child-rights/
[4] https://blogs.lse.ac.uk/medialse/2021/02/04/childrens-rights-apply-in-the-digital-world/

A demo was created in Figma[5] design tool which showed several scenarios in which cyberbullying was detected proactively by AI (i.e., abusive content is detected without the child having to report it to the platform first) on Tik Tok, Instagram and Trill,[6] and subsequent interventions based on research into bystander involvement (Bastiaensens et al., 2014; DiFranzo et al., 2018). For example, a user who witnessed bullying or who had previously been selected by the abused child as a support contact/helper would be automatically notified when their friend is abused and prompted to provide help by reaching out to the victim to offer help; or by being prompted to report to the platform or reach out to the perpetrator soliciting them to stop the abusive behaviour. Another intervention leveraged facial recognition in order to detect bullying by deliberate exclusion (e.g. when three girls want to show the fourth one that she has not been invited to a party by tagging her in their photo from the party, which was previously identified by Instagram as a common type of bullying among teen girls on the platform).[7] The study participants were asked about their perceptions of desirability of AI detecting cyberbullying in this manner, including the use of facial recognition; implications for privacy and freedom of expression, as well as the perceptions of effectiveness of such interventions.

Method and data analyses

This study relies on qualitative research with pre-teen and teen children aged 12-17 (15 semi-structured in-depth interviews conducted online, 8 females, 7 males) and 4 focus groups/FGs (4 groups with female participants conducted offline, in a school setting, and 2 online FGs with males, with 6-10 children per group) from Ireland. All fieldwork was conducted from May to August 2021 and all but the 4 school-based FGs were conducted online due to lockdown conditions. Children were shown the demo with proposed interventions and asked to provide feedback and suggest changes to the interventions. All procedures were approved by Dublin City University's Research Ethics Committee (REC) as well as the Data Protection Unit, parental/caregiver written consent as well as child written assent were sought from all participants following the provision of plain language statements which explained the voluntary nature of the research, confidentiality and anonymity and all research implications in a child friendly manner. Following the transcription and anonymisation procedures, a thematic analysis of the data (Boyatzis, 1998; Braun & Clarke, 2006) was undertaken by three coders, who discussed the themes that emerged and any disagreements as to how the content was coded.

Results and discussion

Despite privacy concerns, many of the interviewed children in both interviews and focus groups approved, would not mind or would even encourage proactive AI "scanning" or "monitoring" on content shared publicly, if it is for the purpose of "greater good" such as preventing abuse. Some children said they were reluctant to allow this on content

[5] https://www.figma.com/
[6] https://www.trillproject.com/
[7] https://about.fb.com/news/2019/05/2019-global-safety-well-being-summit/

shared privately (e.g. direct messages/DMs) but a number of them would allow it under the same rationale, and some children assumed such monitoring is undertaken by platforms already. Some children held the view that whatever one posts online (even in DMs) is probably monitored by platforms or governments anyway. Many children, however, appeared not to be aware of facial recognition and found it “creepy,” but they would point out that they would welcome the use of facial recognition for bullying detection. While children overall liked the idea of having the option to add a support contact/helper/friend who could be notified automatically when AI detected their abuse, many expressed doubts about the effectiveness of the proposed subsequent interventions. Some said they would not necessarily use this support for the fear of overwhelming their support contacts; stigma around asking for help (“one deals with one’s problems on their own”); and they would not want others to know they have a helper/friend/support contact. We discuss these findings in the context of children’s rights and their effectiveness from the perspective of research into bystander support and online safety education (Finkelhor et al., 2021).

References

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders’ behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, *31*, 259-271.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. sage.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Gillespie, T. (2017). Governance of and by platforms. *SAGE handbook of social media*, 254-278.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 2053951719897945.

DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018, April). Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Finkelhor, D., Walsh, K., Jones, L., Mitchell, K., & Collier, A. (2021). Youth internet safety education: aligning programs with the evidence base. *Trauma, violence, & abuse*, *22*(5), 1233-1247.

Livingstone, S., & Third, A. (2017). Children and young people’s rights in the digital age: An emerging agenda. *New media & society*, *19*(5), 657-670.

Mascheroni, G., & Siibak, A. (2021). *Datafied Childhoods: Data Practices and Imaginaries in Children's Lives*. Peter Lang.

Staksrud, E. (2016). *Children in the online world: Risk, regulation, rights*. Routledge.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one, 15*(12), e0243300.

Funding acknowledgment: Tijana Milosevic received support to execute this research from Marie Skłodowska-Curie (MSCA) fellowship[8] and Facebook/Meta Content Policy Award. We thank youth organisation Foróige for assistance with participant recruitment.

[8] https://marie-sklodowska-curie-actions.ec.europa.eu/