Selected Papers of #AoIR2023:
The 24th Annual Conference of the
Association of Internet Researchers
Philadelphia, PA, USA / 18-21 Oct 2023

DESIGNING ETHICAL ARTIFICIAL INTELLIGENCE (AI) SYSTEMS WITH MEANINGFUL YOUTH PARTICIPATION: IMPLICATIONS AND CONSIDERATIONS

Kanishk Verma
ADAPT Centre, Anti Bullying Centre, Dublin City University

Brian Davis
ADAPT Centre, Dublin City University

Tijana Milosevic
Anti Bullying Centre, Dublin City University

James O'Higgins Norman
Anti Bullying Centre, Dublin City University

Introduction

Social media platforms such as TikTok, Snapchat, Instagram, and Facebook offer a range of features for children and young people to interact with one another. The advancements in Internet communication technologies (ICTs), artificial intelligence (AI), and augmented reality (AR) have enabled such platforms to provide many features, including digital effects or *"filters"* that enhance photos and videos. Supported by AI algorithms that detect faces and other objects, such filters help beautify the media content further. Moreover, AI-enabled content recommendation and personalisation allow platforms to keep users engaged for long periods. Social media platforms also leverage AI-based content moderation or proactive moderation to filter out harmful online content before showing it to users. Recent transparency reports by Instagram, Facebook, TikTok, and YouTube indicate an upward trend in proactive moderation of online content followed by a decline in user-reported online content [1]–[3]. While promising, there is limited insight into how these AI algorithms function in real-life scenarios, the user behaviour patterns they learn, and the potential risks like content filtering failures, privacy breaches, and restrictions on freedom of expression.

Recent studies suggest that children and young people are concerned about AI failures and biases in surveillance and profiling by AI [4], [5]. While it is vital to involve children

Suggested Citation (APA): Verma, K., Davis, B., Milosevic, T., O'Higgins Norma, J. (2023, October). Designing Ethical Artificial Intelligence (Ai) Systems With Meaningful Youth Participation: Implications And Considerations. Paper presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from http://spir.aoir.org.

and young people in crafting AI content filters, ethical concerns persist regarding informed consent, participation, mitigation of biases (both implicit and explicit) in AI, and ensuring safeguards for their well-being throughout the AI design [6], [7]. Additionally, safeguarding children and young participants from re-exposure to potential traumatic content is paramount. This article attempts to shed light on ethical concerns in AI system design and possible steps for mitigating those concerns.

AI algorithmic opaqueness

AI techniques leveraged by social media platforms to combat harmful online content remain somewhat opaque. However, insights from strides in computational research offer potential strategies utilised by such platforms [8]. Competitions like those by Semantic Evaluation[1] (SemEval) [9]–[14], have advanced the development of systems capable of identifying various online harms. Despite such systems demonstrating high accuracy in detecting harmful content, they are not near-perfect and often lack consideration for children's viewpoints in their design. Notably, studies focusing on explainable filtering of online bullying text content have made strides in computational research [15], [16], but fall short in incorporating children's perspectives to enhance the transparency of the AI filtering system's decision-making process.

Current ethical risks in youth participation in AI design

AI system development relies heavily on extensive datasets for training and validation, necessitating the annotation of a vast amount of data. Devising such systems for identifying online harms hinges on meticulously annotated datasets. These labelled datasets are crucial for training algorithms to understand the complexities of online content and improve their efficiency in recognising various forms of harm. By engaging children and young people in role-playing online bullying and aggression scenarios, computational researchers [17], [18], were able to devise conversations that serve as training data for AI systems to identify both bullying and aggression. Although the role-playing technique effectively mirrors real-life scenarios and facilitates a more comprehensive representation of bullying and aggression conversations for AI to discern harmful content, it presents numerous challenges as well. One significant limitation is the potential risk of re-traumatising children who might have previously faced similar situations, potentially causing emotional distress when they engage in such design activities.

However, excluding children and young people from the design process can result in differing perspectives on identifying harmful and non-harmful content. A research study by [19], found notable disparities between annotations by domain experts in [17], and

[1] https://semeval.github.io ; Semantic Evaluation is a recurring series of computational semantic analysis competitions currently in its 17th edition. In this competition, the goal is to test computational systems to see how well they can understand and analyse language. The competition has multiple challenges organised for different language processing, generation or inference tasks. The organisers of different tasks in the competition provide text for the computer programs to analyse and also provide the labels or annotations for the text.

children. Thereby implying that children and young people have higher thresholds for defining online content as harmful than domain experts.

Ethical considerations going forward

While building resilience towards online harm is one of the key motivations to involve children and young people in the annotations and design of AI systems, the risk of re-traumatisation from an ethical standpoint is quite severe. Apart from obtaining informed consent from children and their parents or guardians before being involved in such research, there should be other risk mitigation strategies. Engaging in design sessions with children and young individuals, as exemplified by prior studies [20]–[22], involves conducting extensive workshops over several weeks with a diverse research team. Ensuring the research team includes trained professionals, such as psychologists or counsellors to offer support to children during the research process is essential. Additionally, recent findings by [19], highlight the effectiveness of annotation of online harmful content via “interactive games” for children serves as a platform to (a) educate them on sensitive topics, (b) differentiate between harmful and benign content, and (c) gather pertinent annotation through innovative design approaches. Granting children and young people control over their participation and the right to withdraw if uncomfortable or distressed is paramount and places their well-being above research objectives. Prioritising participants’ safety might involve considering alternative methodologies, such as utilising and augmenting previously collected relevant data. Adopting a user-centred and age-appropriate approach remains critical, supported by both pre-and post-design surveys to gauge participant sentiments and well-being. Additional ethical considerations encompass establishing comparison groups, including other annotators and non-participating youth, to evaluate experiences and well-being during sensitive data annotation. Moreover, implementing a continuous feedback (“human-in-the-loop”) mechanism is crucial for ethical AI system design. This approach not only encourages user engagement but also ensures that the AI system remains adaptable and responsive to evolving user needs. Furthermore, facilitating focus-group participation in structured sessions provides a better platform for open dialogue, enabling participants to express their perspectives, concerns, and ideas regarding AI system design. Ethically addressing implicit and explicit biases within AI systems requires deliberate efforts to recognise, analyse, and mitigate these biases. Diverse representation becomes pivotal during system design, necessitating the involvement of children from varied ethnic, cultural, and economic backgrounds. This involvement ensures that the AI system accounts for a broad spectrum of perspectives, minimising biases that might otherwise lead to unfairness or discrimination.

Conclusion

In conclusion, recent studies underscore the concerns of children and young individuals regarding biases and failures in AI surveillance and profiling. Engaging them in the

design of content-filtering AI systems is crucial. However, ethical and rights-based concerns persist regarding informed consent, participation, and addressing biases, necessitating safeguarding measures for children in AI system design. To effectively tackle these challenges, it is crucial to not only secure informed consent from both children and their guardians but also conduct pre-study and post-study surveys. These surveys serve as invaluable tools to assess the well-being and sentiments of the participants throughout the research process. The research team's composition, inclusive of trained professionals like psychologists or counsellors, is essential to provide the necessary support to children facing distress during the research process. Furthermore, conducting workshops and design thinking sessions involving groups of children and young individuals rather than one-on-one interactions in AI system design research is recommended. Implementing a human-in-the-loop AI system stands as a critical measure to counteract potential biases replicated by AI systems in real-world scenarios. These measures collectively contribute to a more ethical and responsible approach to involving children and youth in AI system development.

References

[1] Google, 'YouTube Community Guidelines enforcement', Google Transparency Report. [Online]. Available: https://transparencyreport.google.com/youtube-policy/flags

[2] Meta, 'Community Standards Enforcement | Transparency Center'. Accessed: Oct. 26, 2023. [Online]. Available: https://transparency.fb.com/reports/community-standards-enforcement/bullying-and-harassment/facebook/

[3] TikTok, 'Community Guidelines Enforcement Report'. [Online]. Available: https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2023-2/

[4] V. Charisi *et al.*, 'Artificial Intelligence and the Rights of the Child : Towards an Integrated Agenda for Research and Policy', JRC Publications Repository. Accessed: Oct. 01, 2022. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/handle/JRC127564

[5] 'Examining artificial intelligence technologies through the lens of children's rights'. Accessed: Feb. 24, 2023. [Online]. Available: https://joint-research-centre.ec.europa.eu/jrc-news/examining-artificial-intelligence-technologies-through-lens-childrens-rights-2022-06-22_en

[6] S. Kotilainen, 'Methods in practice: Studying children and youth online', 2022, doi: 10.21241/SSOAR.83031.

[7] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, 'The perspective of European children'.

[8] Kanishk Verma, Brian Davis, Tijana Milosevic, 'Examining the effectiveness of Artificial Intelligence (AI) based cyberbullying moderation on online platforms: Transparency Implications', in *AoIR 2022: Children 2: Safety and Policy*, Dublin, Ireland, Nov. 2022.

[9] V. Basile *et al.*, 'SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter', in *Proceedings of the 13th International*

Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 54–63. doi: 10.18653/v1/S19-2007.
[10] M. Zampieri *et al.*, 'SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)', in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1425–1447. doi: 10.18653/v1/2020.semeval-1.188.
[11] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, 'SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense', in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 105–119. doi: 10.18653/v1/2021.semeval-1.9.
[12] E. Fersini *et al.*, 'SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 533–549. doi: 10.18653/v1/2022.semeval-1.74.
[13] C. Perez-Almendros, L. Espinosa-Anke, and S. Schockaert, 'SemEval-2022 Task 4: Patronizing and Condescending Language Detection', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 298–307. doi: 10.18653/v1/2022.semeval-1.38.
[14] 'SemEval 2023 Task 10: Explainable Detection of Online Sexism (EDOS) | ACL Member Portal'. Accessed: Feb. 24, 2023. [Online]. Available: https://www.aclweb.org/portal/content/semeval-2023-task-10-explainable-detection-online-sexism-edos
[15] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan, 'Does BERT Pay Attention to Cyberbullying?', in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '21. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 1900–1904. doi: 10.1145/3404835.3463029.
[16] K. Verma, T. Milosevic, and B. Davis, 'Can Attention-based Transformers Explain or Interpret Cyberbullying Detection?', in *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 16–29. Accessed: Feb. 24, 2023. [Online]. Available: https://aclanthology.org/2022.trac-1.3
[17] R. Sprugnoli, S. Menini, S. Tonelli, F. Oncini, and E. Piras, 'Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying', in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 51–59. doi: 10.18653/v1/W18-5107.
[18] A. Ollagnier, E. Cabrio, S. Villata, and C. Blaya, 'CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game', in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 867–875. Accessed: Feb. 24, 2023. [Online]. Available: https://aclanthology.org/2022.lrec-1.91
[19] F. Bonetti and S. Tonelli, 'An Analysis of Abusive Language Data Collected through a Game with a Purpose', in *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation*

Conference, Marseille, France: European Language Resources Association, Jun. 2022, pp. 1–6. Accessed: Feb. 24, 2023. [Online]. Available: https://aclanthology.org/2022.games-1.1

[20] K. G. Camelford and C. Ebrahim, 'The Cyberbullying Virus: A Psychoeducational Intervention to Define and Discuss Cyberbullying Among High School Females', *J. Creat. Ment. Health*, vol. 11, no. 3–4, pp. 458–468, Oct. 2016, doi: 10.1080/15401383.2016.1183545.

[21] Z. Ashktorab and J. Vitak, 'Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers', in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, in CHI '16. New York, NY, USA: Association for Computing Machinery, May 2016, pp. 3895–3905. doi: 10.1145/2858036.2858548.

[22] A. N. M. Leung, N. Wong, and J. M. Farver, 'You Are What You Read: The Belief Systems of Cyber-Bystanders on Social Networking Sites', *Front. Psychol.*, vol. 9, 2018, Accessed: Feb. 24, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00365