

A person with glasses is looking at a tablet. The background is a blurred library with bookshelves. The image is overlaid with a blue gradient on the right side.

DCU Anti-Bullying Centre

A background photograph of a school hallway with students. A young woman in a red shirt is in the foreground, smiling and looking towards the camera. Other students are visible in the background, some sitting and some standing. The image has a teal overlay.

Dr. Tijana Milosevic, Elite-S Research Fellow, ABC & ADAPT SFI

Research team: Kanishk Verma, Derek A. Laffan, Dr. Michael Carter, Samantha Vigil, Dr. Brian Davis, Prof. James O'Higgins Norman

DCU Anti-Bullying Centre



Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Social Media Platforms' Efforts at Addressing Cyberbullying among Children

Committee on the Impact of Social Media on the Health Effects of Adolescents and Children

National Academies of Science
February 10th, 2023

Cyberbullying (CB): Definition & characteristics

- Repeated harm inflicted through the use of digital media (Hinduja & Patchin, 2015; Smith & al., 2008)
- Digital iteration of peer-based aggression, public health concern in the US (Espelage et al., 2018)
- Lack of consensus regarding definitions
- Characteristics:
 - Intention to harm
 - Repetitive nature
 - Power imbalance
 - (Gaffney et al., 2019; Centers for Disease Control and Prevention, 2014)



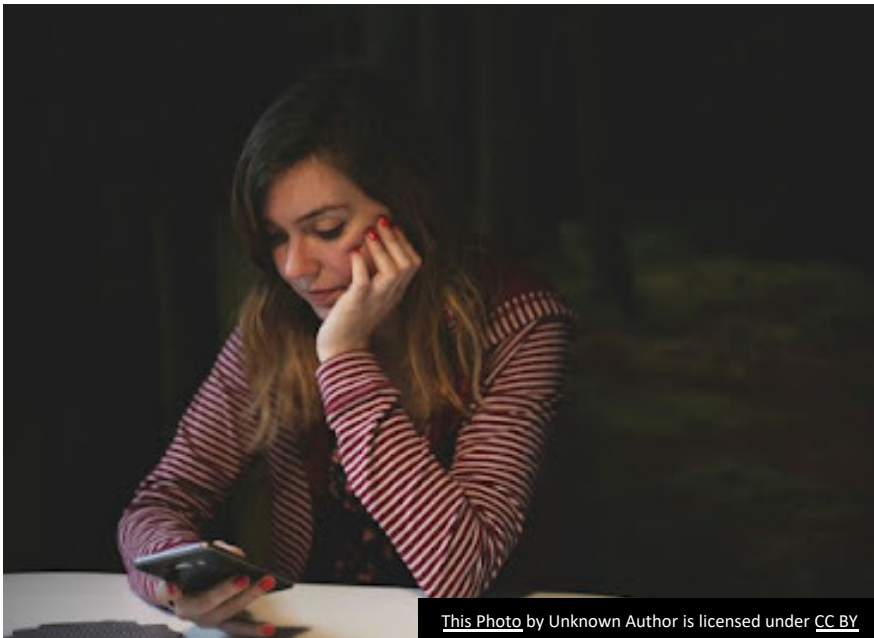
CB prevalence rates

- US, children 13-17:
 - 40% increase since the start of the pandemic
 - 23% targeted in the past month

(Cyberbullying Research Center, 2021)

- Europe, Internet-using 9-16 year-olds:
 - 7% at least every month in the past year
 - 11% a few times in the past year

(Smahel et al., 2020/EU Kids Online Project)



Cyberbullying moderation on social media

Social media companies' CB policies

- Against the policy (Terms of Service, Community Guidelines/Standards)
- Often interchangeable with harassment
- More established/mature companies- Safety Centers, more information on CB definitions, examples
- Hesitance to reveal internal moderation documents
- Leaked internal documents: failure to provide adequate protections
(Gillespie, 2018; Milosevic, 2018; Shipp et al., 2022)

PROTECTING CHILDREN ONLINE?

CYBERBULLYING POLICIES
OF SOCIAL MEDIA COMPANIES



TIJANA MILOSEVIC

FOREWORD BY SONIA LIVINGSTONE

How Facebook allows users to post footage of children being bullied

Leaked guidelines on cruel and abusive posts also show how company judges who 'deserves our protection' and who doesn't

- Revealed: Facebook's rules on sex, terrorism and violence
- 'No grey areas': experts call for change

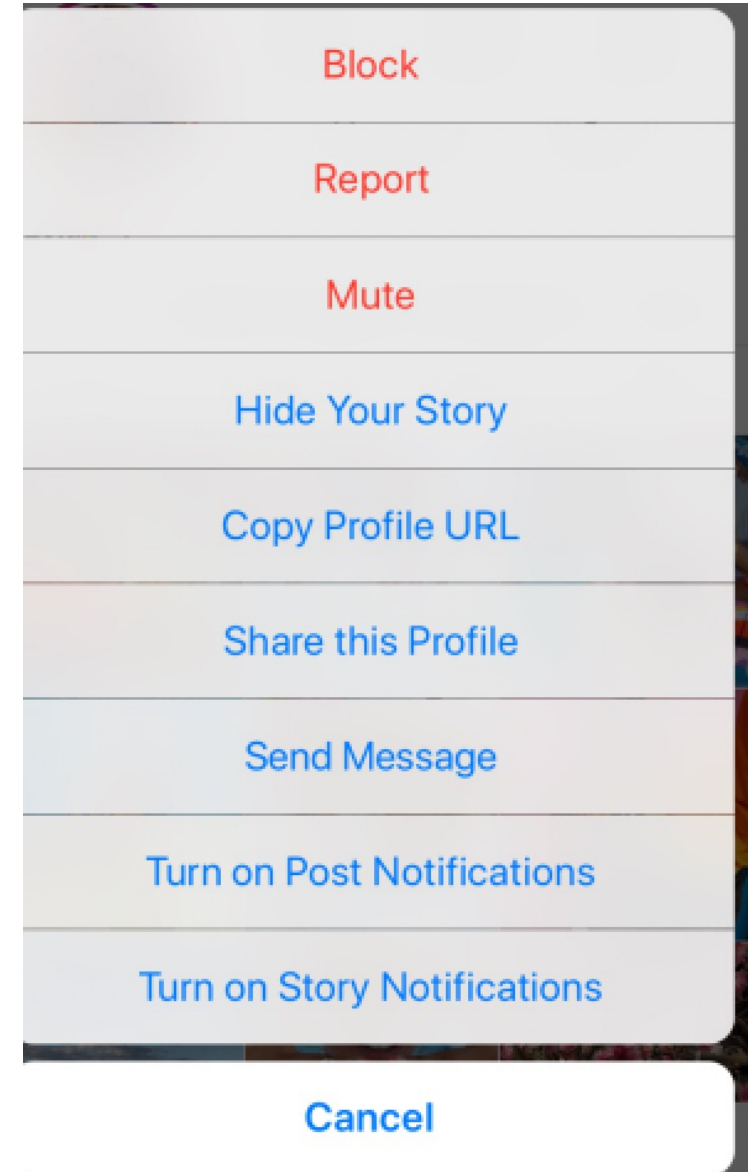


Footage of 'physical bullying' of children under seven is allowed on Facebook as long as there is no caption.
Composite: Alamy

The Guardian, 2017

Enforcement: Reactive moderation

- Reporting, blocking
 - Mute, restrict
 - Comment filtering
 - Earlier efforts (social reporting, Facebook)
 - Escalation (pilot, schools, Facebook)
 - Content removal:
 - Data on effectiveness?
 - Effectiveness from children's perspective?
- (Milosevic & Vladislavljevic, 2020; Roberts, 2019)

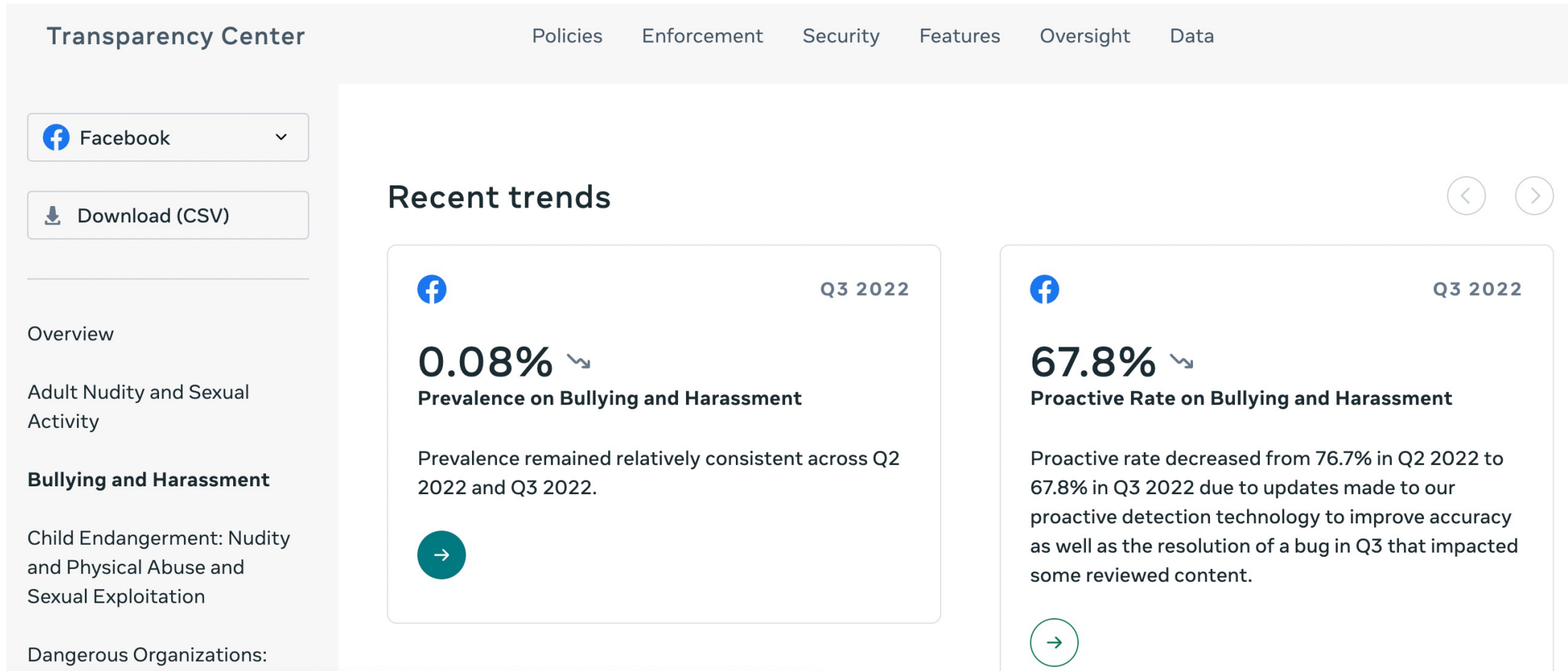


Proactive moderation (Algorithmic, AI-based)

- Technologies used for detection (Gorwa et al., 2021; Kirk et al., 2022; Vidgen & Derczynski, 2020)
- Publicly available resources for independent researchers to scrutinise industry efforts are scarce:
 - Proprietary nature of Facebook AI's DeepText, Linformer, RIO and WPIE: not much can be known about their effectiveness in tackling CB
 - Primarily hate speech detection—not the same as CB
 - Google and Jigsaw's Perspective—effective at tackling toxicity on online platforms, it can still be deceived by subtle modifications to the text.

(Verma et al., 2022)

Proactive moderation, Transparency reports



From: <https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook/>

Regulation

- Digital Services Act (EU)
- Online Safety and Media Regulation Bill (OSMR), Ireland
- Online Safety Bill, UK
- Online Safety Act, Australia
 - Systemic regulation
 - Individual complaints mechanism (e.g. IE, Australia)
 - Online Safety Codes
 - Auditing
 - Safety by design
 - Age-appropriate design (e.g. ICO, UK)
 - Duty of care (UK)
 - Youth Councils (IE, Australia)
- Shortcomings of transparency?
- Transparency for the sake of meaningful auditing:
 - Evaluating effectiveness from children's perspective (i.e. with children/by children)

Importance of engaging children in policy design

- United Nations Convention on the Rights of the Child (UNCRC), 1989
- Rights apply in digital environments (General Comment)
 - Rights to protection
 - But also provision, participation and privacy
- AI based interventions that establish the balance of rights

(Committee on the Rights of the Child, 2021; Livingstone, 2021; Livingstone, Carr, Byrne, 2016)



This Photo by Unknown Author is licensed under [CC BY](#)



Our research

- RQ1: How can we design automatic tools that support effective proactive bullying interventions that assist 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 optimize the effectiveness of such tools and ensure the detection of subtle bullying?

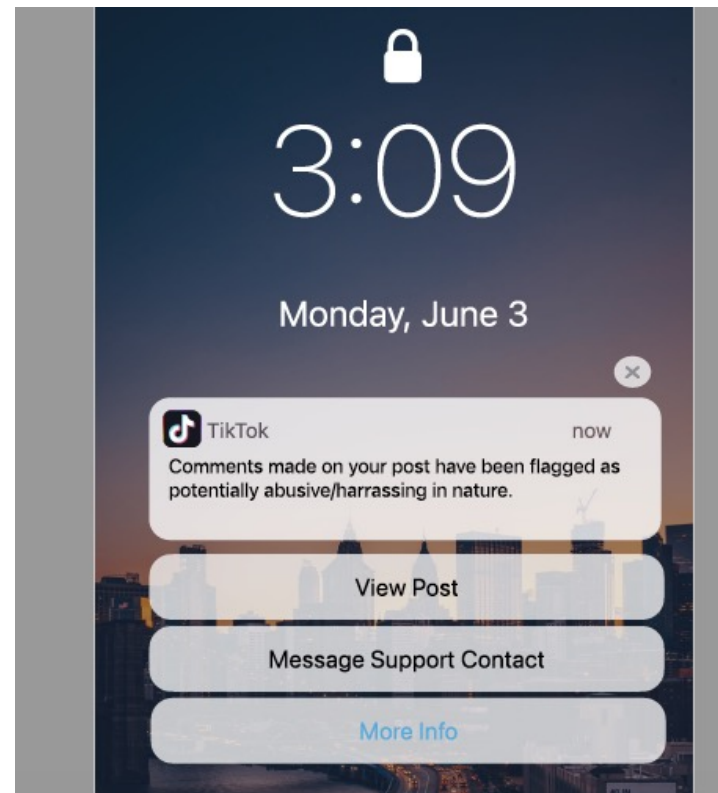
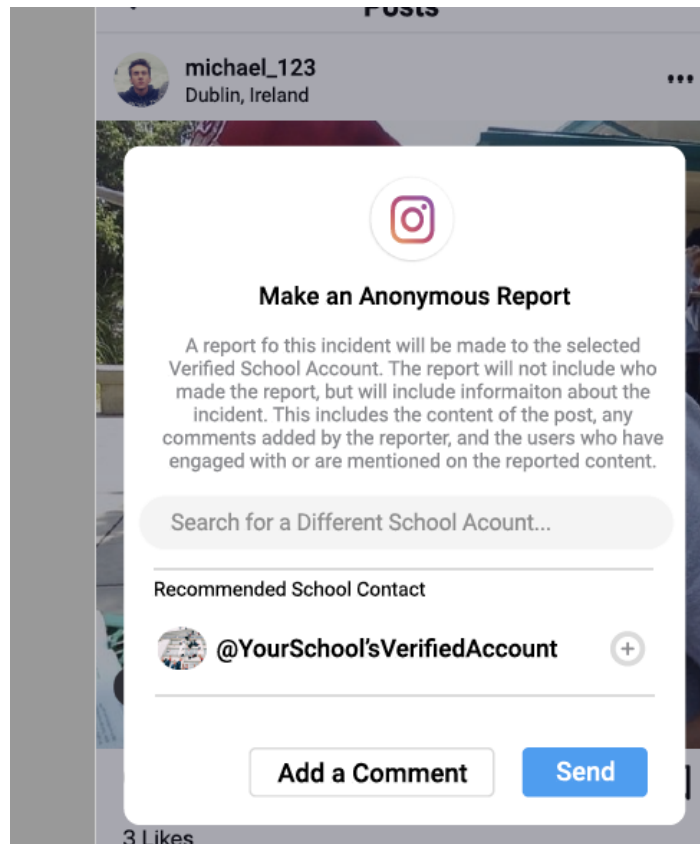
Designing hypothetical interventions

- Peer support (Support Contact/Helper)
- Bystander involvement
 - Finding the right way to engage bystanders
 - Accountability
 - Rendering witnessing process visible
- AI-triggered school involvement
- Support scores and unlocking platform features
- Less post engagement for perpetrators for a limited amount of time

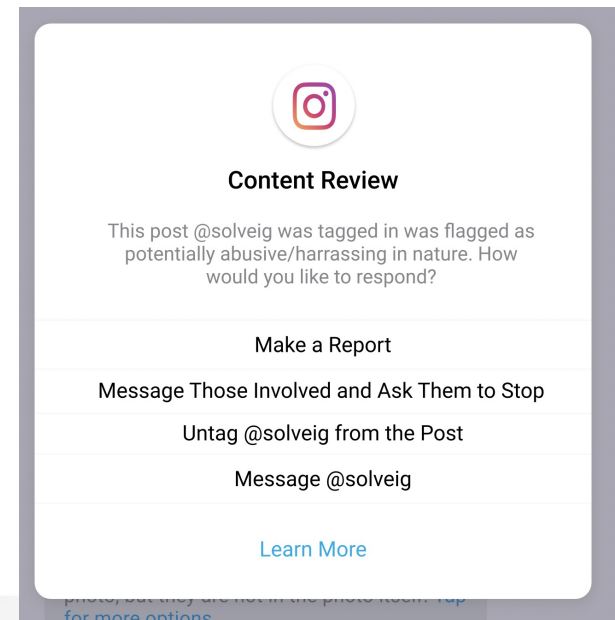
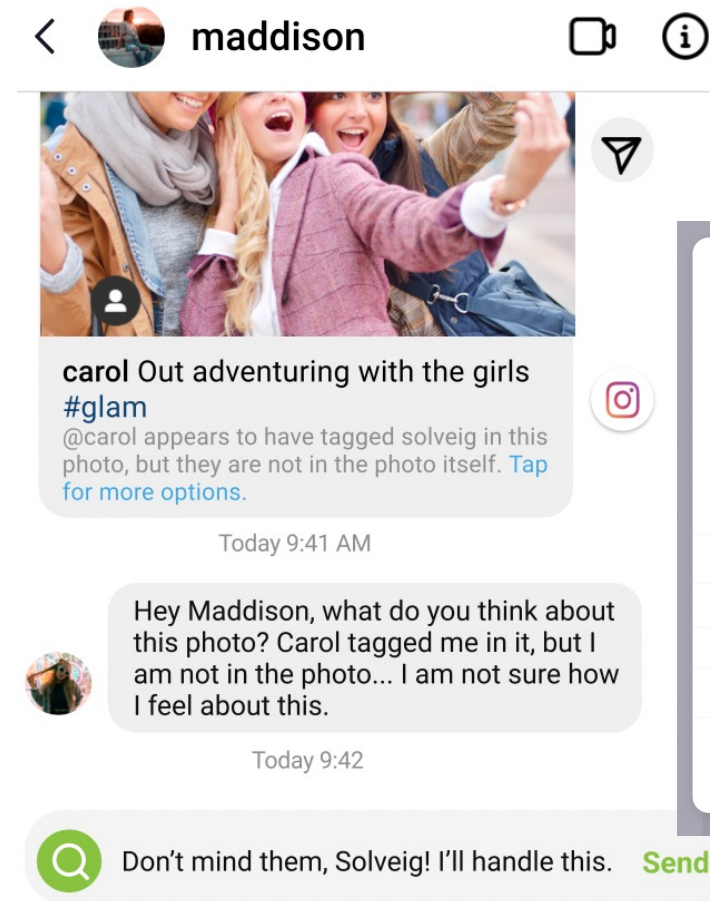
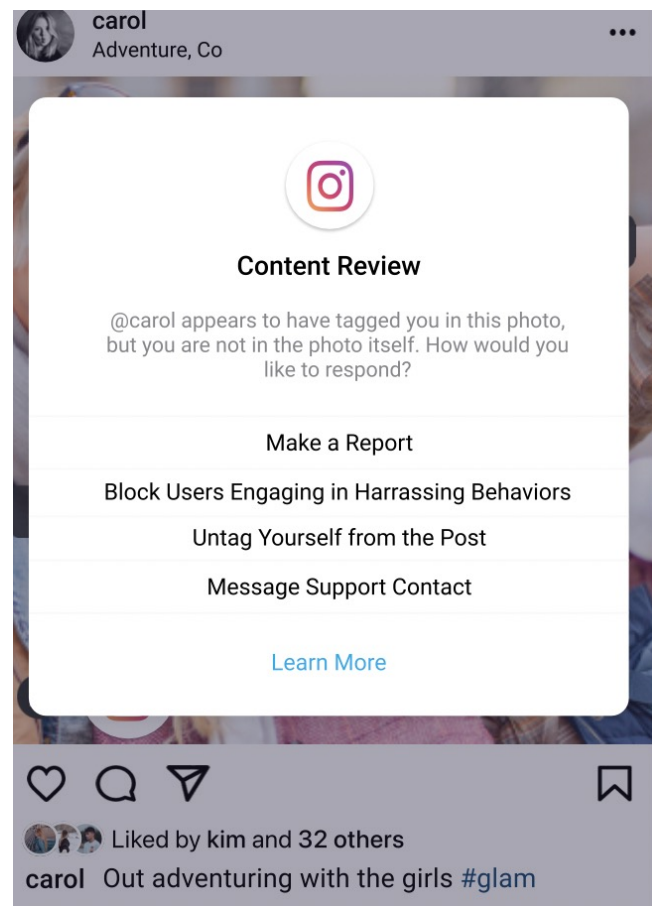
(Bastiaensens et al., 2014; Bowler, Knobel & Mattern, 2015; DiFranzo et al., 2018; Milosevic, 2018; Mishna et al., 2021; Van Royen et al., 2017)

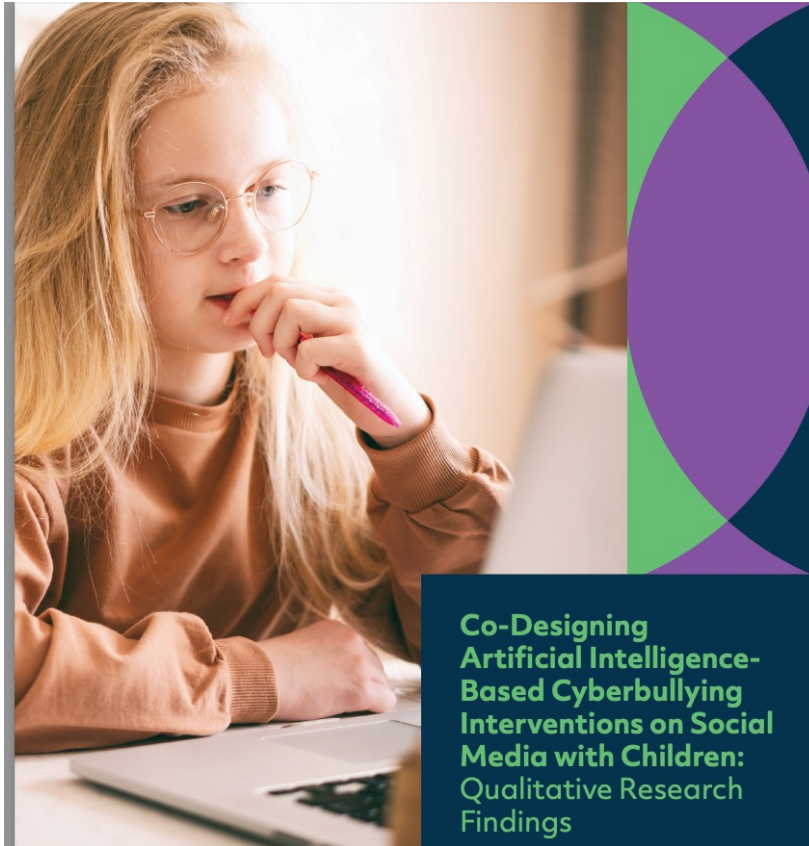
Method

- Qualitative research
- Children age 12-17
- 4 focus groups with teen girls (at school, in person)
- 2 focus groups with teen boys (Zoom)
- 15 individual in-depth interviews (Zoom)
- June –August 2021



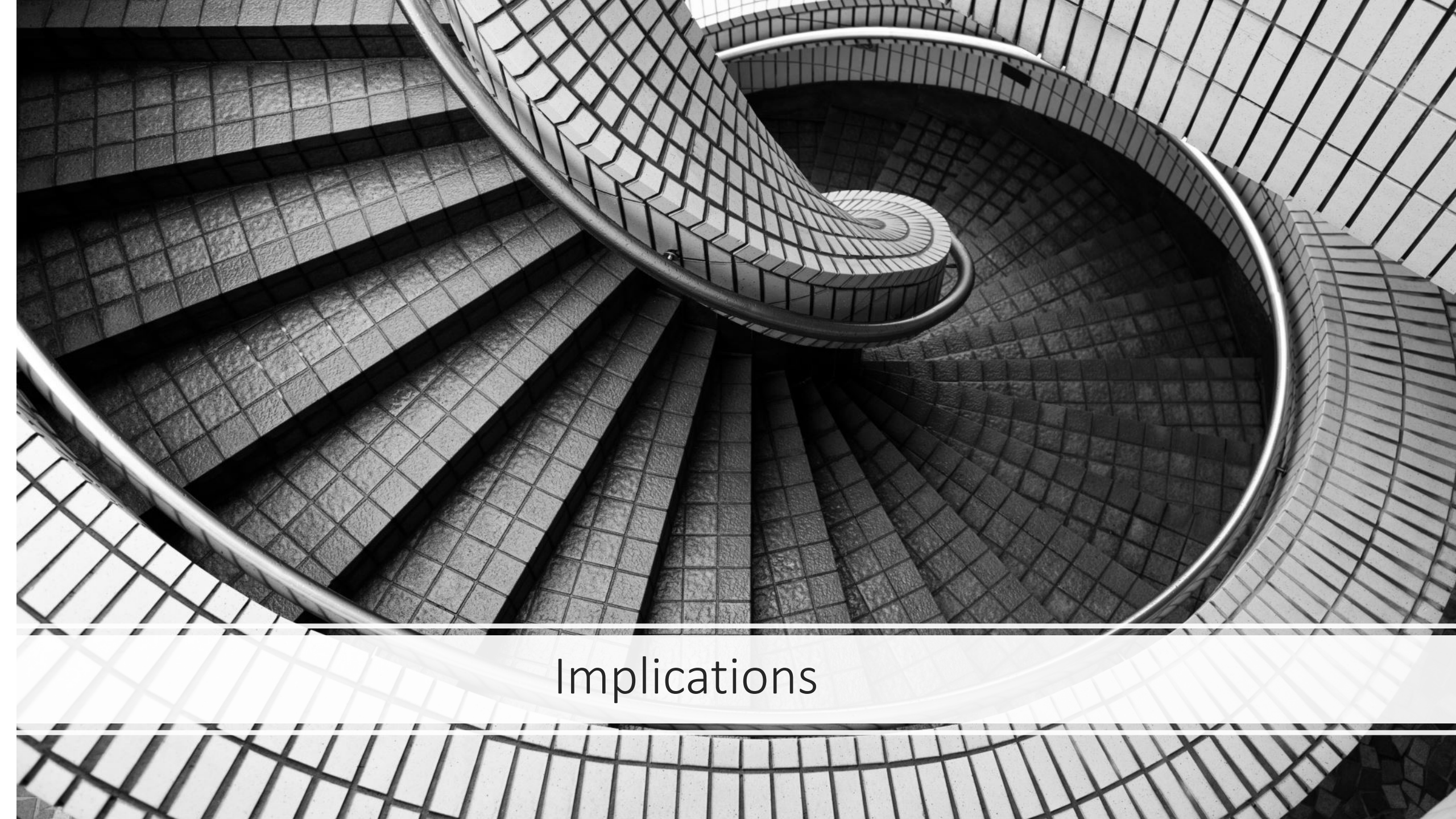
Figma-based demos (scenarios)





Results

- AI overall welcomed, but not sure if they would use them
- Privacy concerns, direct messaging and facial recognition
 - Perception of “ok if for greater good”
- Clear opt in and out
- Issues around admitting one has a support contact, burdening the support contact
- Hesitancy around bystander involvement
- Social norms: self-reliance, asking for help for sensitive children
- Some interventions perceived as more suitable for younger children
- Full findings here:
<https://journals.sagepub.com/doi/full/10.1177/20563051221147325>
- (Milosevic et al., 2023)



Implications



Thank you

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. Committee on the Rights of the Child. (2021).
- Bowler, L., Knobel, C., & Mattern, E. (2015). From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology*, 66(6), 1274-1293.
- Cyberbullying Research Center. (2021). Cyberbullying Statistics, 2021. Retrieved from: <https://cyberbullying.org/cyberbullying-statistics-age-gender-sexual-orientation-race>
- 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).
- Espelage, D. L., Hong, J. S., & Valido, A. (2018). Cyberbullying in the United States. *International Perspectives on Cyberbullying: Prevalence, Risk Factors and Interventions*, 65-99.
- Gaffney, H., Farrington, D. P., Espelage, D. L., & Ttofi, M. M. (2019). Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review. *Aggression and violent behavior*, 45, 134-153.
- General Comment on the Rights of the Child in Relation to the Digital Environment. Retrieved from: <https://www.ohchr.org/EN/HRBodies/CRC/Pages/GCChildrensRightsRelationDigitalEnvironment.aspx>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- 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.
- Hicks, D. (2018). *Leading with dignity: How to create a culture that brings out the best in people*. Yale University Press.
- Hinduja, S., & Patchin, J. W. (2015). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin press.

References

- Kirk, H. R., Vidgen, B., & Hale, S. A. (2022). Is More Data Better? Re-thinking the Importance of Efficiency in Abusive Language Detection with Transformers-Based Active Learning. *arXiv preprint arXiv:2209.10193*.
- Lindner, E. (2012). A dignity economy: Creating an economy that serves human dignity and preserves our planet. Oslo: Dignity Press.
- Livingstone, S., Carr, J., & Byrne, J. (2016). One in three: Internet governance and children's rights. Retrieved from: <http://disde.minedu.gob.pe/handle/20.500.12799/4686>
- Livingstone, S. (2021, March 24th). Children's rights apply in the digital world! Retrieved from: <https://blogs.lse.ac.uk/parenting4digitalfuture/2021/03/24/general-comment-25/>
- Milosevic, T. (2018). Protecting children online?: Cyberbullying policies of social media companies. The MIT Press.
- Mishna, F., Birze, A., Greenblatt, A., & Khoury-Kassabri, M. (2021). Benchmarks and bellwethers in cyberbullying: the relational process of telling. *International Journal of Bullying Prevention*, 3(4), 241-252.
- Milosevic, T., Verma, K., Carter, M., Vigil, S., Laffan, D., Davis, B., & O'Higgins Norman, J. (2023). Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From Youth Perspective. *Social Media+ Society*, 9(1), 20563051221147325.
- National Advisory Council for Online Safety. (2021). Report of a National Survey of Children, their Parents and Adults Regarding Online Safety.
- Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., ... & Hasebrink, U. (2020). EU Kids Online 2020: Survey results from 19 countries.
- Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4), 376–385.
- Shipp, J., Mitchell, A., Noula, I., Grady, P. (2022). Accountability Report 2.0. Internet Commission. Retrieved from: <https://inetco.org/>
- Roberts, S. T. (2019). Behind the screen. Yale University Press.
- Van Royen, K., Poels, K., Vandebosch, H., & Adam, P. (2017). "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior*, 66, 345-352.



Ionad
Frithbhulaíochta
Anti-Bullying
Centre

