Co-Designing Artificial Intelligence-Based Cyberbullying Interventions on Social Media with Children: Qualitative Research Findings

By: Tijana Milosevic, Kanishk Verma, Samantha Vigil, Michael Carter, Derek Laffan, Brian Davis, & James O'Higgins Norman

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Abstract

This report details the results of a qualitative research study (focus groups and in-depth interviews) with children and teens aged 12-17 (N=59) in Ireland about the perceived effectiveness of Artificial Intelligence (AI)-based cyberbullying enforcement mechanisms on popular social media platforms. The adoption of the UN General Comment No. 25 established that children's rights, as outlined in the UN Convention on the Rights of the Child (UNCRC), apply in a digital environment. We therefore examine children's perceptions about how AI-based enforcement mechanisms affect their rights to protection (safety), participation and privacy. We inquire into how children perceive the effectiveness of the proposed mechanisms; and how these could be made more effective from their perspective; and which changes or alternatives they propose. The proposed interventions are based on social learning and social norm theories, and they include designated support contacts, bystander and school involvement, and systems that are designed to reward prosocial behaviours and deter perpetration. We find that children would welcome many interventions but raise concerns around their privacy and effectiveness of what has been proposed. We provide policy recommendations for the technology industry and policy makers.

Tijana Milosevic, Elite-S Research fellow, DCU Anti-bullying Centre (ABC) and ADAPT SFI

Kanishk Verma, Irish Research Council PhD Candidate, DCU School of Computing, ADAPT SFI, ABC DCU

Samantha Vigil, PhD Student, Department of Communication, University of California, Davis

Michael Carter, PhD Candidate (ABD), Department of Communication, University of California, Davis

Derek Laffan, ABC DCU

Brian Davis, Professor, DCU School of Computing and ADAPT SFI

James O'Higgins Norman, Director ABC, DCU, Professor at DCU, and UNESCO Chair on Tackling Bullying in Schools and Cyberspace

Introduction

During Covid-19 lockdowns, youth overwhelmingly relied on the Internet for activities that normally take place offline, such as schooling. While also for socialising and leisure, in some countries, this uptick in mediated activities coincided with an increased rate of cyberbullying victimisation for particular age groups (Lobe et al., 2021). Cyberbullying, or the enactment of repeated and intentionally hurtful behaviour (Hinduja & Patchin, 2015), is a serious problem across social media platforms and can take on various forms. For example, cyberbullying can span mean or abusive comments, posts or direct messages (DMs); creating a fake profile of someone for the sake of mocking them; excluding someone from an activity on purpose, revealing their private information (eg, doxing); and so on (Smith, 2016; O'Higgins Norman, 2020). Given its complexity, cyberbullying definitions[1] remain a matter of academic and policy debate, and cyberbullying is sometimes considered as interchangeable with harassment, especially in social media platforms' policies. Nevertheless, platforms typically do not allow activities deemed as abuse, cyberbullying, and/or harassment on their platforms, as stipulated in their Terms of Service, Community Standards, Guidelines, or other comparable documentation (Milosevic, 2016, 2018).

Common mechanisms implemented over popular social media platforms as tools to target instances of cyberbullying also vary in their level of human involvement. Conventionally, users often have access to options for reporting on abusive content uploaded or sent over a platform. This typically initiates a moderation process to determine whether reported content or activities violate the company's policy and if any content should be taken down. With millions of users and vast amounts of content, it is impossible for companies to rely on human moderators alone to facilitate this process, however (Gillespie, 2018). Algorithmic applications, such as natural language processing (NLP), machine learning (ML) and deep learning (DL), remain common among popular apps to help automate the process of content moderation; these approaches fall under the umbrella of "artificial intelligence" or AI (Gorwa et al., 2020; Milosevic et al., 2022). Given the capacity of AI-based moderation techniques, companies have begun to use AI applications to try and proactively moderate content by detecting and removing content before it is even reported by users (Community Standards Enforcement Report).[2] Notwithstanding such innovations, the legitimacy and efficacy of the use of AI for content moderation over social media remains under ongoing scrutiny (Heldt & Dreyer, 2021). In the end, platforms often have to partially rely on users to take the initiative, whether by reporting content or by implementing various forms of user blocking (eg, unfriending, blocking), content restriction (eg, segregating audiences, muting), and/or content filtering (eg, Hidden Words on Instagram) to manage their experiences on the platform.[3]

[1] UNESCO and the World Anti-Bullying Forum. (November 1-3, 2022). Presenting a proposed revised definition of school bullying. Retrieved from: https://delegia-virtual.s3.eu-north-1.amazonaws.com/projects/delegia-wabf/WABF_summary_of_new_definition.pdf

[2] Meta (2021, November 9). Community Standards Enforcement Report: Third Quarter 2021. Retrieved from: https://about.fb.com/news/2021/11/community-standards-enforcement-report-q3-2021/

[3] Instagram Help Centre (2022). How do I filter our and hide comments I don't want to appear on my posts on Instagram? Retrieved from: https://help.instagram.com/700284123459336

As a result of ongoing developments in platform-based practices for addressing instances of cyberbullying and the persistent reliance of platforms on user involvement to intermediate their experiences with forms of online abuse, it is critical to understand youths' perspectives on the efficacy of platform tools in this context. This is especially true given the rise of legislative frameworks focusing on systemic changes to content circulation (for an example, see Douek, 2022), which require the provision of evidenced effectiveness of platform tools and AI-based moderation internationally (eg, Online Safety and Media Regulation Bill[4], Ireland; Online Safety Bill[5], the United Kingdom; Digital Services Act,[6] the European Union; Online Safety Act,[7] Australia), which were in part created to protect young users. Lastly, provided the pace at which the commercial social media landscape changes over time, there remains ample need to explore innovative and novel platform mechanisms targeting the mitigation of online forms of cyberbullying and abuse in particular.

Therefore, the present study sought to advance understanding of youths' perspectives towards AI mechanisms and platform tools targeting cyberbullying and online forms of abuse over multiple social media apps. To do so, we conducted 6 focus groups and 15 semi-structured in-depth interviews with children and adolescents (N =59) to assess their views towards platform mechanisms targeting online abuse through a set of five hypothetical cyberbullying scenarios illustrated via a set of realistic, yet mock user interfaces mirroring core aspects of commercially available social media apps (i.e., TikTok, Instagram, Trill Project). Scenarios included an array of novel platform mechanisms (eg, designated support contact, bystander notifications, and rewards) to explore a range of possible intervention designs. The study represented the qualitative phase of the research project "Co-designing with Children: A rights-based Approach to Fighting Bullying" funded by Facebook/Meta Content Policy Award, Phase 2.[8] In all, results help to inform the developing policy-making environment globally in the context of social media by highlighting themes in youths' perspectives towards different types of platform mechanisms targeting instances of cyberbullying and abuse online.

[4] Government of Ireland. (2022, January 25). Publication of the Online Safety and Media Regulation Bill. Retrieved from: https://www.gov.ie/en/publication/88404-publication-of-the-online-safety-and-media-regulation-bill/

[5] Gov. UK, Department for Digital, Culture and Sport. (2022, March 17). Online Safety Bill: FactSheet. Retrieved from: https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-factsheet

[6] European Commission.(2022, March 25). The Digital Services Act Package. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[7] Australian Government (n.d.). Federal Register of Legislation: Online Safety Act 2021. Retrieved from: https://www.legislation.gov.au/Details/C2021A00076

[8] Meta. (2019). Announcing the winners of phase two content policy research awards. Retrieved from: https://research.facebook.com/blog/2019/09/announcing-the-winners-of-phase-two-content-policy-research-awards/

Children's rights in digital environments: implications for safety by design[9]

The adoption of the United Nations' General Comment No. 25[10] in 2021 established that children's rights as specified in the United Nations Convention on the Rights of the Child (UNCRC) apply in the digital world (Staksrud, 2016; Livingstone et al., 2016). This signifies that children have, among others, rights to **protection**, and freedom from abuse and cyberbullying is considered as a right to be protected and safe online and offline. They also have the right to **provision**, which encompasses the right to education and quality media content, for example. Bullying and cyberbullying in schools are considered as an affront to education because they interfere with the child's ability to learn. Therefore, ensuring the right to protection from bullying and cyberbullying is a critical enabler of other rights. Finally, children also have the right to **participation** which includes their right to express views on matters that concern them (such as cyberbullying moderation on social media platforms); and also to participate in environments that provide them with leisure and socialisation opportunities, such as social media.

Another important right that directly concerns this study is children's right to **privacy**. In online environments, children's right to privacy can be violated and infringed upon in several ways (Livingstone et al., 2019). Firstly, children can unwittingly share too much information about themselves which can jeopardise their safety. For example, in more extreme cases, they can publicly reveal information about where they live or leave geolocation traces which can allow for their tracking by strangers. Children's privacy can be jeopardised in the social context too, for example when their friends or parents/caregivers take their photos without their consent and then post them on social media, a phenomenon known as sharenting (Livingstone et al., 2020). Data collection for commercial purposes such as tracking, which is the basis of social media platforms' business models, can also constitute an affront to children's privacy (Mascheroni & Siibak, 2021). Among other issues, this includes sharing their data with third parties and data brokers that can lead to data breaches and profiling that can hamper their future education and job opportunities (Montgomery et al., 2017). In this study, we are interested in children's perceptions of privacy in the context of deploying AI-based cyberbullying interventions. Specifically, we are focused on hearing about how children perceive their right to privacy from AI-based monitoring of public posts and private messages, as well as the use of facial recognition.

[9] Australian Government, eSafety Commissioner. (n.d.). Safety by design. Retrieved from: https://www.esafety.gov.au/industry/safety-by-design

[10] United Nations Human Rights Office of the High Commissioner (2021). General Comment No. 25 (2021) on children's rights in relation to the digital environment. Retrieved from: https://www.ohchr.org/EN/HRBodies/CRC/Pages/GCChildrensRightsRelationDigitalEnvironment.aspx

Protection vs. participation?

In addition to reporting abusive content for company moderation, social media platforms tend to provide users with automated options designed to assist with various forms of cyberbullying where the platform does not get involved. For example, users can block and mute others or discussions that they find insulting; they can restrict users if they do not wish to see their comments and posts (a polite way of blocking), they can also turn off comments so that no one can comment on their content; or they can turn on comment filtering ("Hidden Words"[11] on Instagram), whereby all comments which are detected to have abusive content are not shown/visible to the target.

Such features, however, place the onus on the young user to deal with problems on their own without the platform's assistance. Secondly, the features that restrict activity or access, prioritise children's safety over their ability to fully participate in online environments (Livingstone & Third, 2017). For example, having an Instagram account that is private or if comments are switched off for safety reasons can limit the possibilities of engagement, thus prioritising children's right to protection over their right to participation, as we detail below. Companies often recognise that their support systems face significant challenges and are not fully adequate; some companies also have a trusted flagger system (also known as trusted reporter) in place, which allows individuals and various third-party organisations such as non-governmental organisations the option to draw companies' attention to individual cases of abuse; and escalate cyberbullying, hate speech and other violations to companies.[12]

[11] Instagram Help Centre (2022). How do I filter our and hide comments I don't want to appear on my posts on Instagram? Retrieved from: https://help.instagram.com/700284123459336

[12] YouTube Help. (n.d.). YouTube trusted flagger program. Retrieved from: https://support.google.com/youtube/answer/7554338?hl=en; European Commission (2019). Code of Conducting on Countering Illegal Hate Speech Online. Retrieved from: https://ec.europa.eu/info/sites/default/files/code_of_conduct_factsheet_5_web.pdf

Theory and research informing the design of our interventions

Interventions tested in this study (please see section "Interventions tested in this study," which details each intervention) are informed by social learning and social norm theories, which posit that maladaptive behaviours such as cyberbullying are reinforced by role models who behave in an overtly or covertly aggressive manner; and when these behaviours are supported by the social environment or considered as acceptable or normative (Espelage et al., 2012; Hinduja & Patchin, 2013). For example, when the perpetrator receives tacit support or active encouragement from those who witness the abuse, such behaviour enables the perpetrator to continue with abusive behaviour, and it also sends a message to those who witness the abuse (bystanders) that such behaviour is allowed. Much research has therefore focused on the conditions that determine whether a bystander will get involved to assist the victim in a cyberbullying situation (therefore becoming an "upstander").

Furthermore, offline and online bullying tend to go hand in hand, with an overlap between victimisation and perpetration (eg, bully-victim phenomenon); those who were once victims can also be perpetrators at the same time or later on (eg, getting back at someone; Kowalski et al., 2014; Görzig, A., & Macháčková, 2015). Thus, while online the perpetrator can operate anonymously by hiding behind a username or a fake profile/account, young people often know who bullies them, especially if an incident happens in the context of peer relationships or at school (Mishna et al., 2009; Mishna et al., 2021; O'Higgins Norman, 2020).

Existing research has explored social and technology design-related conditions that increase the likelihood of bystander involvement to assist the victim rather than the perpetrator (Bastiaensens et al., 2016; DeSmet et al., 2014). It has been examined whether a lack of empathy and accountability contributed to failure to get involved to help the victim, with some findings suggesting that empathy could prevent negative bystander behaviour (Barlińska et al., 2013; Macháčková & Pfetsch, 2016). As for accountability, it has long been established in research on offline bullying that the presence of more bystanders can diffuse the sense of responsibility whereby each of them believes that someone else will help the victim (Latane & Darley, 1970). Accountability, or the belief that one will be held responsible for one's actions leads to a sense of personal responsibility which motivates action to help the victim, and research has recently explored how technological design can promote a sense of personal responsibility (van Bommel et al., 2012). For instance, it has been found that a sense of awareness that one is being watched in public and that they are not anonymous was found to be conducive to prosocial behaviours (Pfattheicher, & Keller, 2012). For example, when bystanders were informed about the size of the audience and when they received a notification from the platform that they have seen an abusive post/message, they were more likely to intervene on the side of the victim by reporting the bullying content to the platform (DiFranzo et al., 2018). Reporting abusive behaviour, content or blocking the perpetrator is considered as indirect support for the victim, whereas direct support would entail writing to the victim to offer help or responding to attacks or addressing the perpetrator.

Following this line of research, we created a set of demos whereby support from designated contacts/helpers and bystanders is solicited when abusive behaviour such as cyberbullying is detected by AI (as described below). Earlier research conducted by Meta/Facebook in collaboration with Yale Centre for Emotional Intelligence and University of California, Berkeley, explored social reporting, a process which allowed users to solve conflicts amongst themselves by reaching out to others for help with pre-made messages; or to perpetrators with requests to take content down (Anderle, 2016; Milosevic, 2018). The research attempted to test whether users would reach out to third parties for help using pre-made messages and whether this process would result in the perpetrator taking content down or apologising for their actions. According to the findings presented at Facebook's Compassion Research Day, of the 25% of children who used social reporting options, 90% messaged the person they had a problem with, and over a third of those who posted something problematic deleted such content once they had been contacted and asked to do so (Milosevic, 2018, p. 129). Hence, we provided the option in some of our demos for the victim, support contact or bystander to reach out to the perpetrator asking to take the content down or apologise. We also created a variation on this type of a response by allowing children to create an anti-bullying video with pre-made text which tells the perpetrator that such abusive behaviour is hurtful or not ok (variations on the type of message were possible and open to children for feedback).

Reflective messages are a widely researched intervention which was shown to be effective in reducing abusive behaviours, and some platforms already have them in place (Ashktorab & Vitak, 2016; Lieberman et al., 2011; Van Royen et al., 2021; Van Royen et al., 2017; Van Royen et al., 2016). Before posting/messaging, the content of the post is screened by AI for abusive content and if it is detected, the poster/sender is provided with a reflective message prompting them to reconsider if they really wish to post/send it. Since this is a widely researched type of intervention, we used it as an optional demo, and asked participants about desirability and perceived effectiveness of this tool.

Finally, research has suggested that interface design which rewards prosocial behaviours should be explored further in terms of effectiveness in reducing undesired behaviours (Wu et al., 2022). We solicit youth views on the idea to gain access to more platform features and increase one's supportiveness score as a reward for helping others.

The current study

The following research questions guided this phase of the project:

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?

Interventions tested in the study

Interventions we designed in this study involve not only the target (victim) and the perpetrator but also those who witness cyberbullying incidents, the so-called "bystanders" (Rudnicki et al., 2022). Bystanders can remain neutral and not become involved in the incident they are witnessing; or they can support the perpetrator or support the victim (at which point, they are considered to be "upstanders"). Furthermore, we have included a feature called "support contact/helper/friend" whom children can add upon sign up and who can be contacted when abuse is detected by AI. The idea behind the support contact is based on peer mentoring (Papatrainou et al., 2014; Bauman & Yoon, 2014), but we envisaged that the support contact can be an adult as well (parent/caregiver, or someone else who is close to the child).

Using a collaborative interface design tool, Figma,[13] the research team created four core and two optional demos[14] each showing a scenario with an example of abusive behaviour that could constitute a cyberbullying incident on Instagram, TikTok and Trill[15] and a subsequent intervention. Core scenarios were shown in each interview and focus group while the optional ones were shown if there was additional time in the session. Each scenario then showed examples of how the incident could be detected by AI proactively and a subsequent intervention based on research into bystander involvement in cyberbullying incidents (Bastiaensens et al., 2014; DiFranzo et al., 2018; Macaulay et al., 2022). The proposed interventions as designed in this study are hypothetical and only some components of these are available currently on certain social media platforms. For example, "hidden words" on Instagram allow the user to turn on comment filtering, which removes abusive comments which the user can later on nonetheless view if they would like to. All of the features we propose, should be, however, technologically feasible to implement, based on the current state of AI development for the purpose of detecting cyberbullying and harassment as previously identified by the authors of this report (Milosevic et al., 2021).

For example, we proposed that once children create an account on Instagram/TikTok/Trill, children be offered the option to add a support contact/helper/friend who could be contacted if AI detects cyberbullying or some other type of abuse on the platform. A support contact could be a friend, parent, teacher or someone else and the person need not be using the given platform. In Demo 1, (Image sequence 1) we showed an example of a girl receiving negative comments on her post on TikTok; once these are detected by AI, the girl receives a notice

[13] Figma can be accessed here: https://www.figma.com/

[14] All demos can be found on this link: https://drive.google.com/file/d/1O6PzyffWKhjP1SkJedDbgl_qrjFG1HYl/view?usp=sharing

[15] Trill is a social network that allows for anonymous sharing and whose goal is to provide support space for improving mental health: https://www.trillproject.com/

from TikTok that abusive comments have been detected, and she is prompted to review them (abusive comments are not displayed automatically in order not to traumatise her if she chooses not to see them); or to request help from the support contact. Demo 1 also showed the option to request support from those who have been detected by AI as bystanders (eg, they posted something positive or neutral on the post that received negative comments, or have merely been detected as having seen the abusive post). Those identified by AI as bystanders would receive a prompt from the platform that abusive comments have been detected on the person's post and they'd be prompted to intervene by providing support to the person who was abused; or by reporting the abusive content or account to the platform; or by reaching out to the perpetrator asking them to take it down. We then asked children for feedback on the desirability of such options, perceived effectiveness of these interventions and their perceptions of how such deployment of AI might affect their privacy and freedom of expression.

In the second demo, we featured an example of cyberbullying by exclusion, which according to Instagram was a common way for teen girls to experience cyberbullying on the platform.[16] For example, purposeful exclusion would be made visible and performative (Marwick & boyd, 2014) by tagging the person in a story or post featuring photos from the event to which she was not invited. In the demo, we showed three teen girls tagging the fourth one in a photo from an event where she was not invited.

By photo analysis and facial recognition, AI application could detect that more people are tagged in the photo than are actually present in the photo; and establish that bullying has possibly occurred by further examination of direct messages (DMs) exchanged among the three girls who talked about not inviting the fourth one to the event, and then showing her that she was not invited by tagging her in the photos. Thereafter, the victim would receive a prompt asking her whether she'd like to review the post where she'd been tagged in and report it to Instagram, in case it was bullying. Any intervention that would prompt the victim to view an abusive message should contain a trigger warning as well. She would also be prompted to reach out to her support contact for help. The support contact would be provided with the option to reach out to the girls who engaged in exclusion and ask them to take the post/story down, explaining that such behaviour is hurtful. Both the victim and the support contact would have the option to restrict further sharing of this post/story on Instagram and other platforms, in addition to the regular options of reporting it to the platform and untagging themselves.

Demo 3 offered the possibility of reporting a cyberbullying incident on Instagram to one's official school account which would be managed by a professional at their school. Under this scheme, every school in Ireland would have an official account on Instagram. Upon sign up, children would be given an option to confirm their attendance of a particular school and given the ability to report incidents to their school. This demo is a variation on Facebook/Meta's earlier proposals and efforts in the United States (at the state level)[17] to involve schools as escalators or trusted flaggers.

[16] According to information presented at Meta/Facebook Global Safety Summit, 2019: https://about.fb.com/news/2019/05/2019-global-safety-well-being-summit/

[17] The Baltimore Sun. (2013, October 3). Facebook and Md. Schools Partner to Combat Bullying. Retrieved from: https://www.baltimoresun.com/education/bs-xpm-2013-10-03-bs-md-facebook-school-partnership-20131003-story.html

Under such a scheme, the school would be able to flag a case to the platform for prioritised handling as a trusted flagger (Milosevic, 2018). In the demo, we did not position schools as trusted flaggers, but rather we tested the desirability of school involvement into cyberbullying cases altogether. The demo shows a boy tagged in a post with abusive comments underneath; the post was then detected by AI proactively and the boy was prompted to report it to his school in addition to reporting it to the platform; like in previous demos, the option to reach out to a support person was provided; as well as the possibility of asking the perpetrator to take the post down. Furthermore, the perpetrator was punished by having less engagement on all his posts over the course of the following month (i.e., all his posts regardless of the nature of their content would have less visibility to other users on the platform, similarly to shadow banning[18]), following a notification and the option to appeal the decision.

Demo 4 took place on Trill and it showed homophobic bullying of a person via direct messaging. AI was able to scan DMs for abusive content and following the detection of such content, the sender was automatically blocked; and the victim received prompts with options to seek support from the support contact and report the content to the platform. Subsequently, those who engaged as support contacts were rewarded with support score points, which could be added to one's account profile/username and they were also rewarded by being able to unlock additional platform features such as colours.

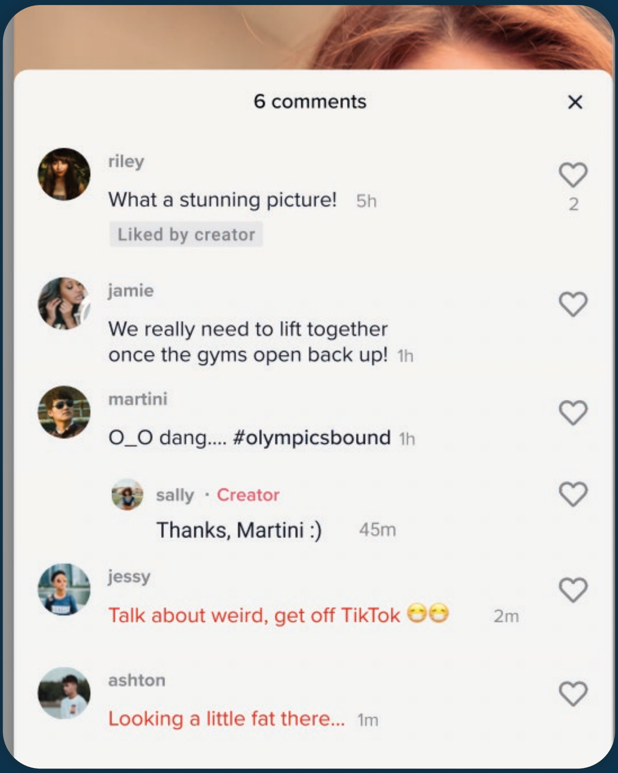
Demo 5 was an optional demo (we only showed it if there was enough time left in the end of each interview/FG session) which allowed users to create an anti-bullying video on TikTok and Instagram upon sign-up. The anti-bullying video could be tailored by the user and created together with the support contact/helper/friend and feature any music/sound clips available. Users could incorporate a pre-made message such as "be kind" or "that was hurtful," or "this is not ok," asking the perpetrator to take abusive content down or stop the abuse (common messages in online safety campaigns[19]); or the user could write something that they thought was appropriate, which could even try to frame the situation in a joking manner or be more assertive in tone towards the perpetrator. The video could then be sent automatically when AI detects something abusive towards the user; or the user could choose whether and when it should be sent.

Finally, the last optional demo showed a "reflective message," a well-researched intervention already used by some platforms, which prompts the user who is about to post something detected as abusive to think twice before posting it. The message that the poster was about to post was not necessarily abusive, it expressed a negative opinion "a bit dull if you ask me" in response to a throwback post of someone having fun in a photo of a pre-Covid lockdown party. The comment was trying to convey the message that their party did not seem like that much fun after all.

[18] TikTok. (n.d.) What is Shadow Banning. Retrieved from: https://www.tiktok.com/discover/what-is-shadow-banning?lang=en

[19] Webwise.ie (n.d.). Be kind online. Retrieved from: https://www.webwise.ie/uncategorized/be-kind-online-sid/; TackleBullying.ie (n.d.). Resources. Retrieved from: https://tacklebullying.ie/resources/

Image sequence 1 (Demo 1): Adding a support contact and AI-triggered request for help



6 comments ✕

riley
What a stunning picture! 5h
Liked by creator
♡ 2

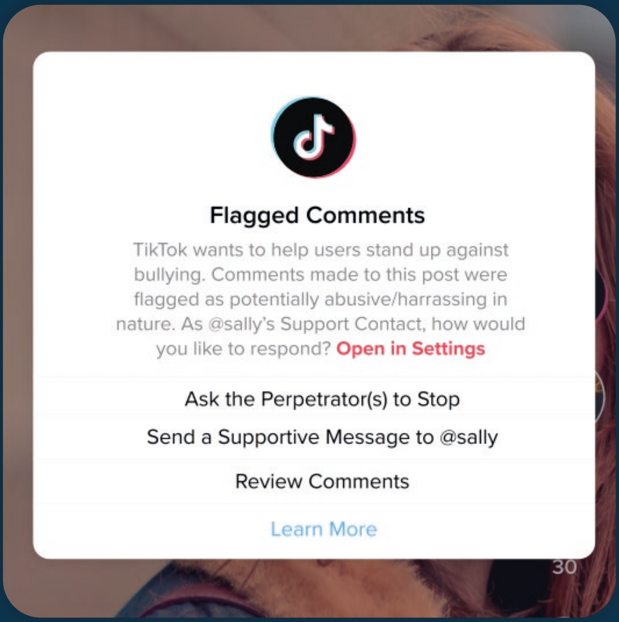
jamie
We really need to lift together once the gyms open back up! 1h
♡

martini
O_O dang.... #olympicsbound 1h
♡

sally · Creator
Thanks, Martini :) 45m
♡

jessy
Talk about weird, get off TikTok 😁😁 2m
♡

ashton
Looking a little fat there... 1m
♡



Flagged Comments

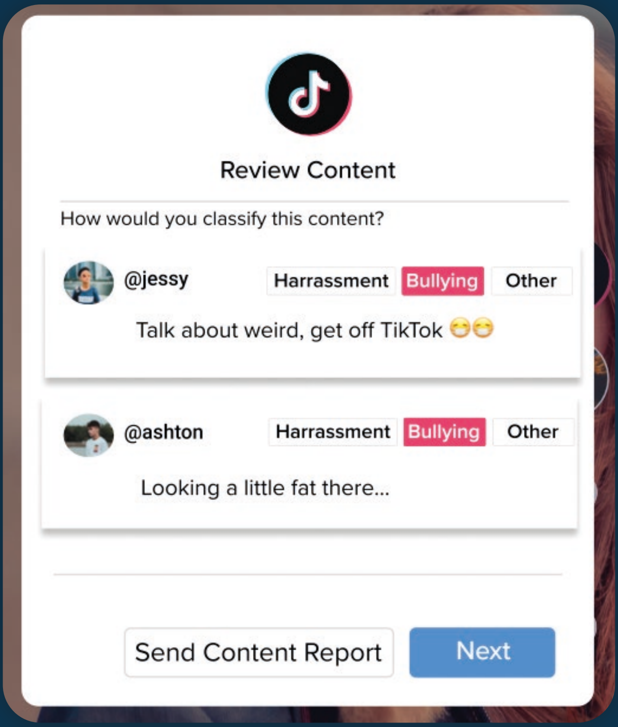
TikTok wants to help users stand up against bullying. Comments made to this post were flagged as potentially abusive/harrassing in nature. As @sally's Support Contact, how would you like to respond? **Open in Settings**

Ask the Perpetrator(s) to Stop

Send a Supportive Message to @sally

Review Comments

Learn More



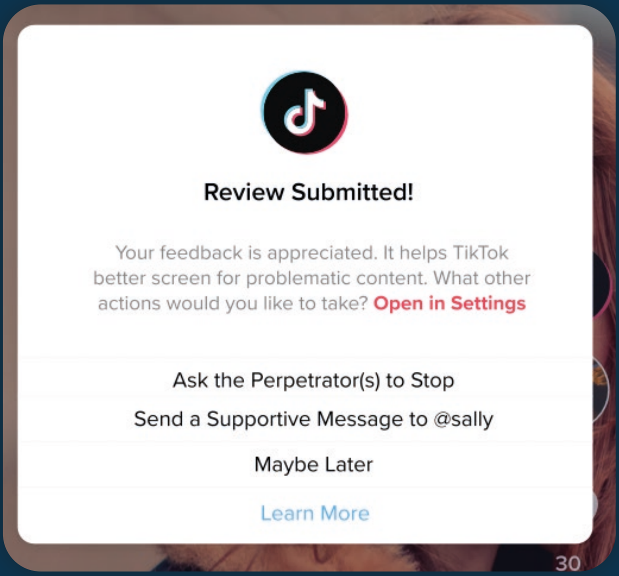
Review Content

How would you classify this content?

@jessy Harrassment **Bullying** Other
Talk about weird, get off TikTok 😁😁

@ashton Harrassment **Bullying** Other
Looking a little fat there...

Send Content Report | Next



Review Submitted!

Your feedback is appreciated. It helps TikTok better screen for problematic content. What other actions would you like to take? **Open in Settings**

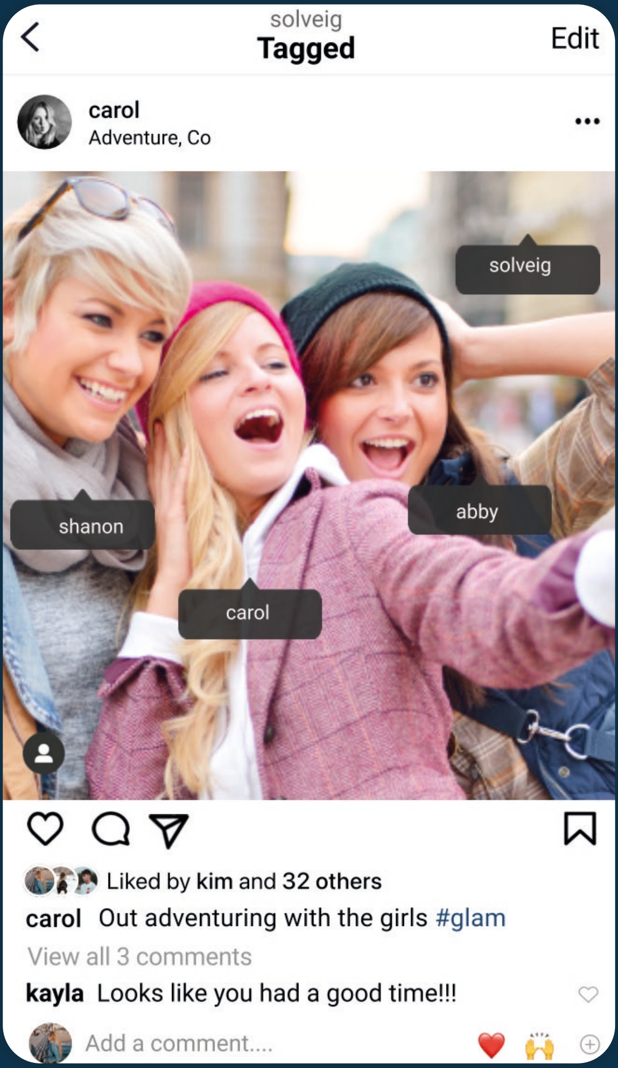
Ask the Perpetrator(s) to Stop

Send a Supportive Message to @sally

Maybe Later

Learn More

Image sequence 2 (Demo 2): Exclusion, AI-based notification



Liked by kim and 32 others

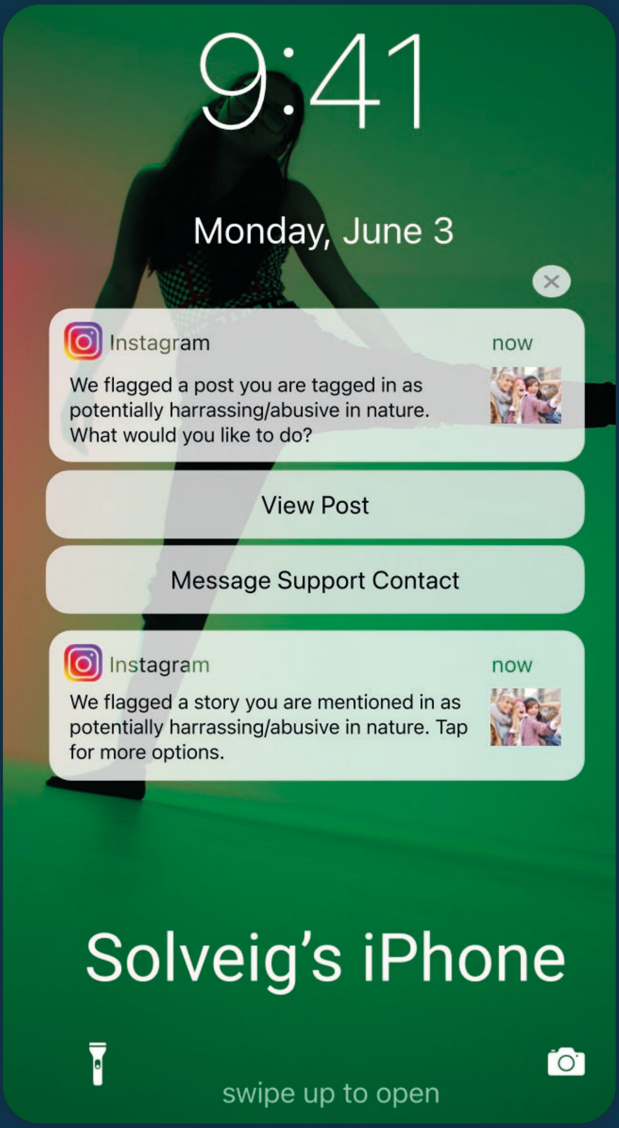
carol Out adventuring with the girls #glam

View all 3 comments

kayla Looks like you had a good time!!!

Add a comment....

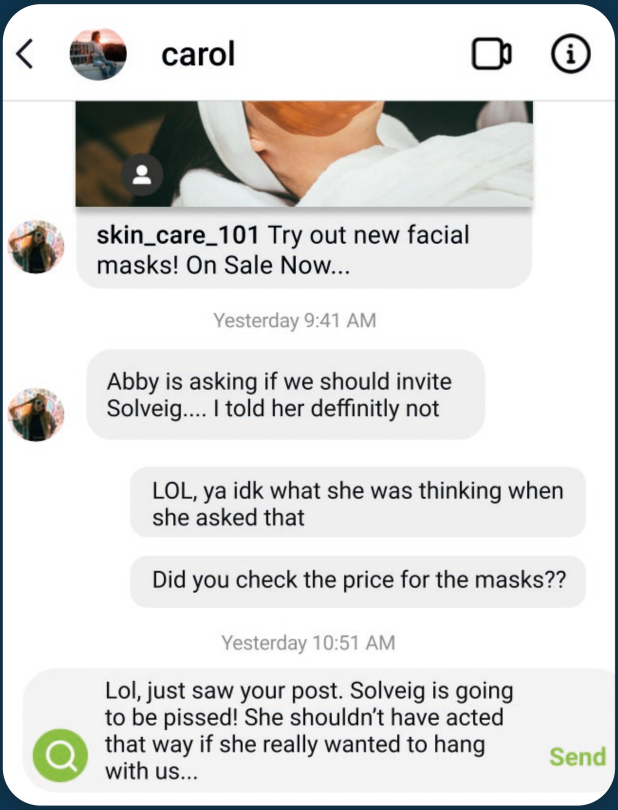
Girls tagging Solveig to show her she is excluded.



We flagged a post you are tagged in as potentially harrassing/abusive in nature. What would you like to do?

We flagged a story you are mentioned in as potentially harrassing/abusive in nature. Tap for more options.

Solveig receiving AI-triggered prompt about the post.



skin_care_101 Try out new facial masks! On Sale Now...

Yesterday 9:41 AM

Abby is asking if we should invite Solveig.... I told her deffinitly not

LOL, ya idk what she was thinking when she asked that

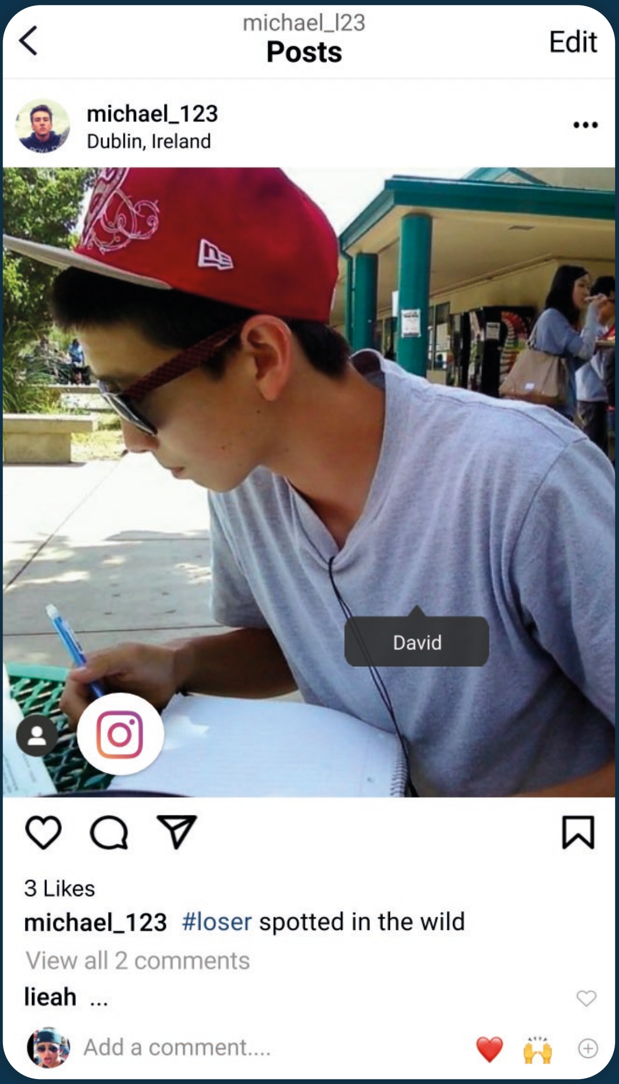
Did you check the price for the masks??

Yesterday 10:51 AM

Lol, just saw your post. Solveig is going to be pissed! She shouldn't have acted that way if she really wanted to hang with us...

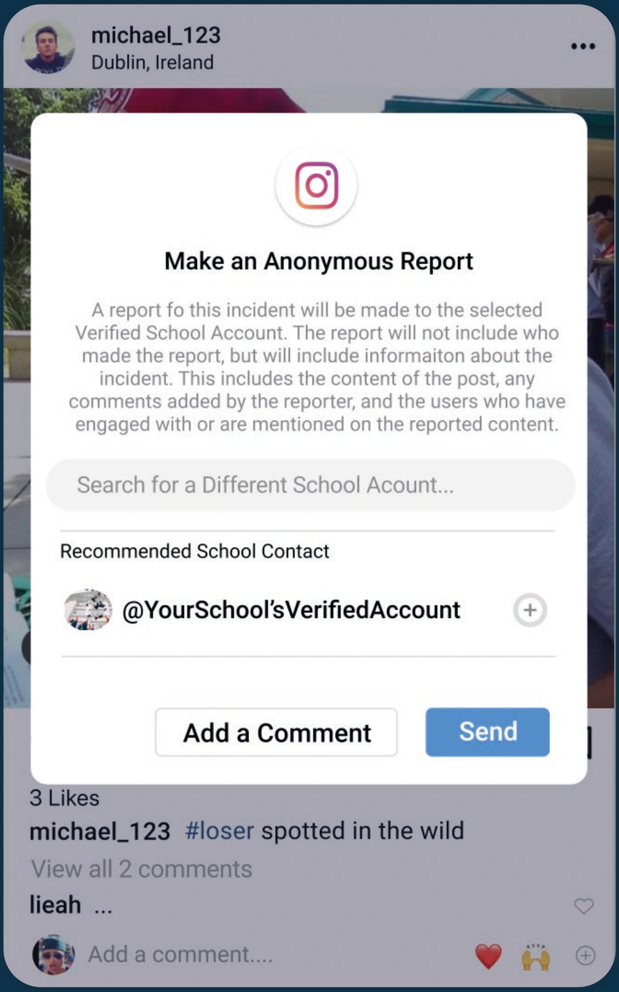
Private chat among the three girls showing they deliberately excluded Solveig.

Image sequence 3 (Demo 3): Reporting to the verified school account



michael_123 #loser spotted in the wild

lieah ...



Make an Anonymous Report

A report fo this incident will be made to the selected Verified School Account. The report will not include who made the report, but will include informaiton about the incident. This includes the content of the post, any comments added by the reporter, and the users who have engaged with or are mentioned on the reported content.

Search for a Different School Acount...

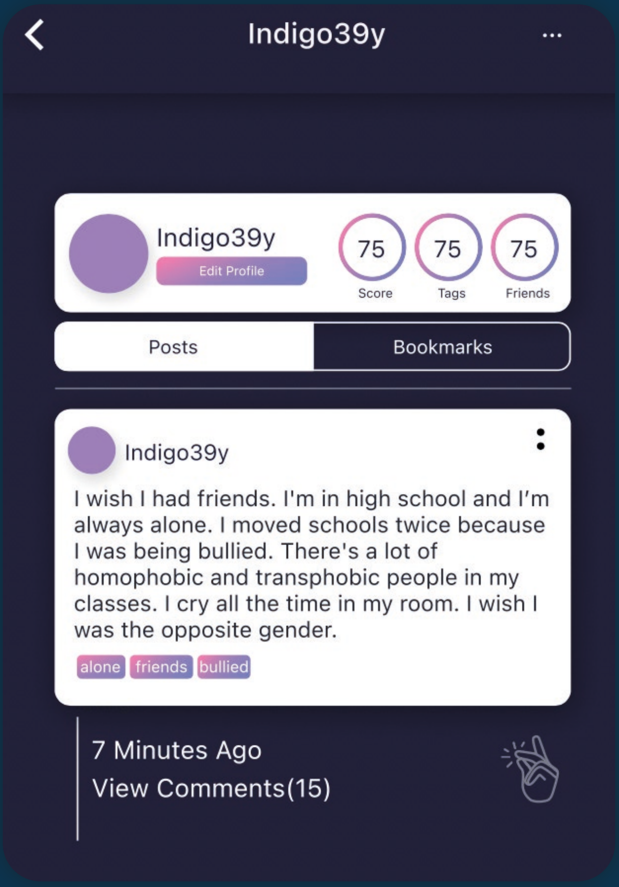
Recommended School Contact

@YourSchool'sVerifiedAccount

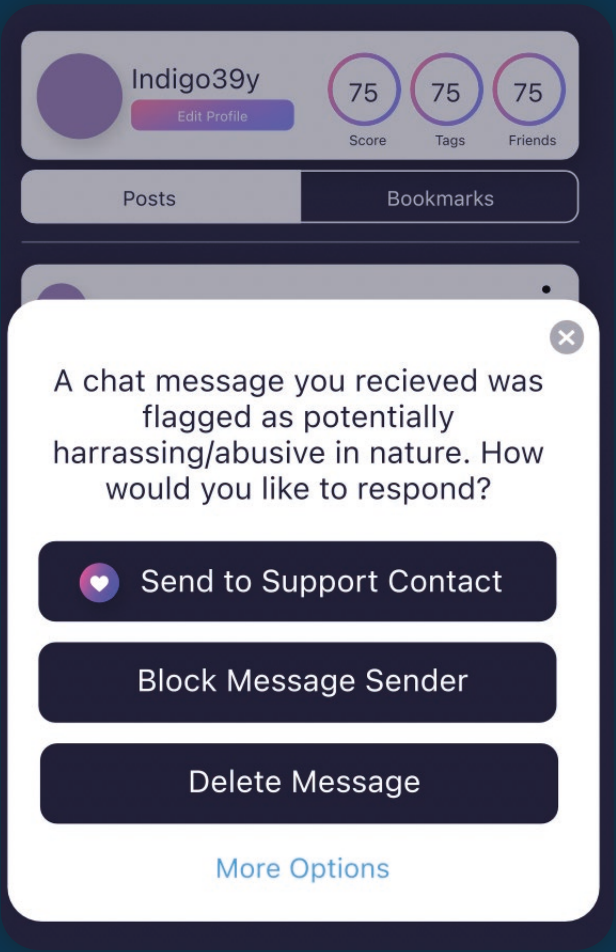
michael_123 #loser spotted in the wild

lieah ...

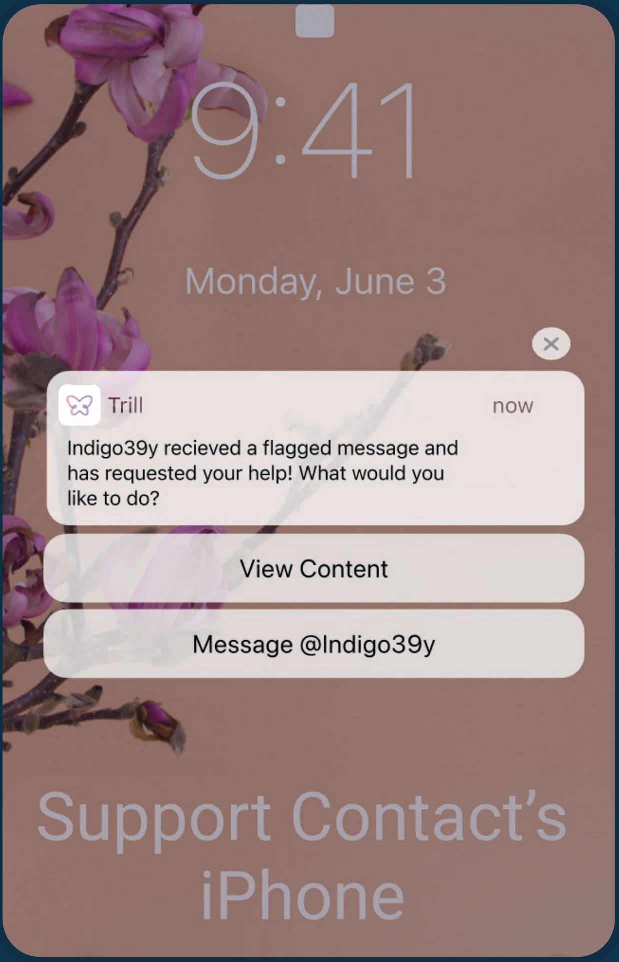
Image sequence 4 (Demo 4): Abusive DM, Requesting help from support contact on Trill



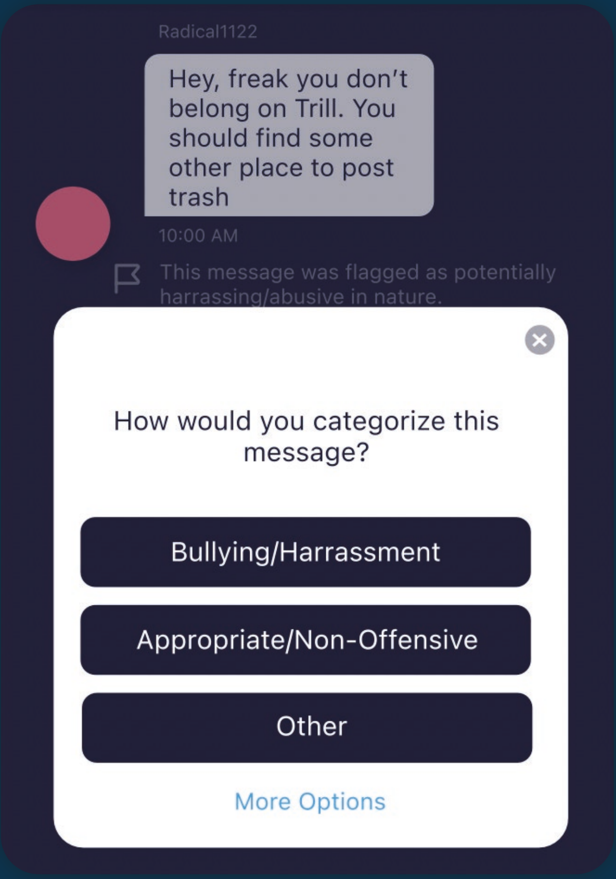
I wish I had friends. I'm in high school and I'm always alone. I moved schools twice because I was being bullied. There's a lot of homophobic and transphobic people in my classes. I cry all the time in my room. I wish I was the opposite gender.



A chat message you recieved was flagged as potentially harrassing/abusive in nature. How would you like to respond?



Indigo39y recieved a flagged message and has requested your help! What would you like to do?

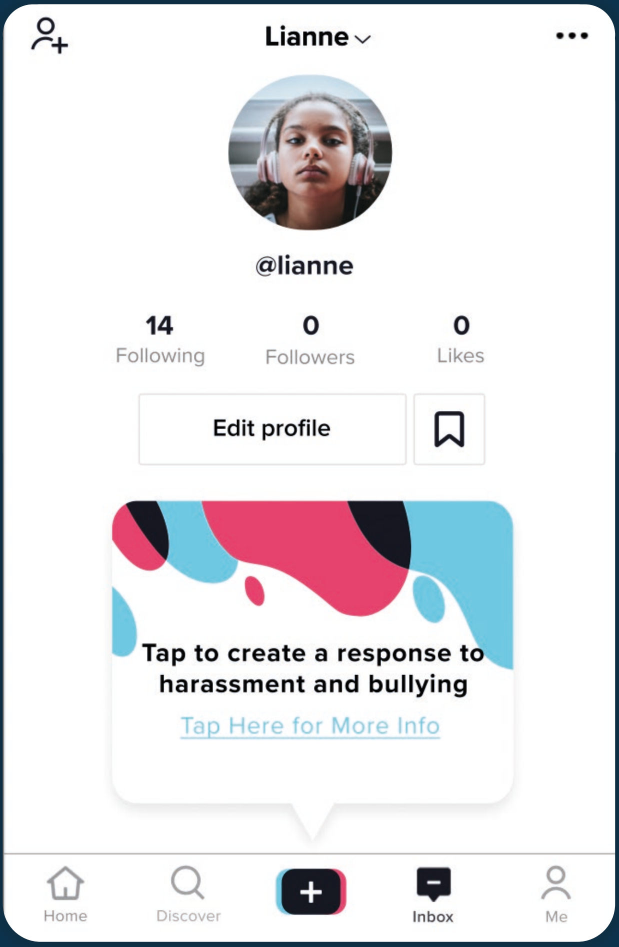


Hey, freak you don't belong on Trill. You should find some other place to post trash

This message was flagged as potentially harrassing/abusive in nature.

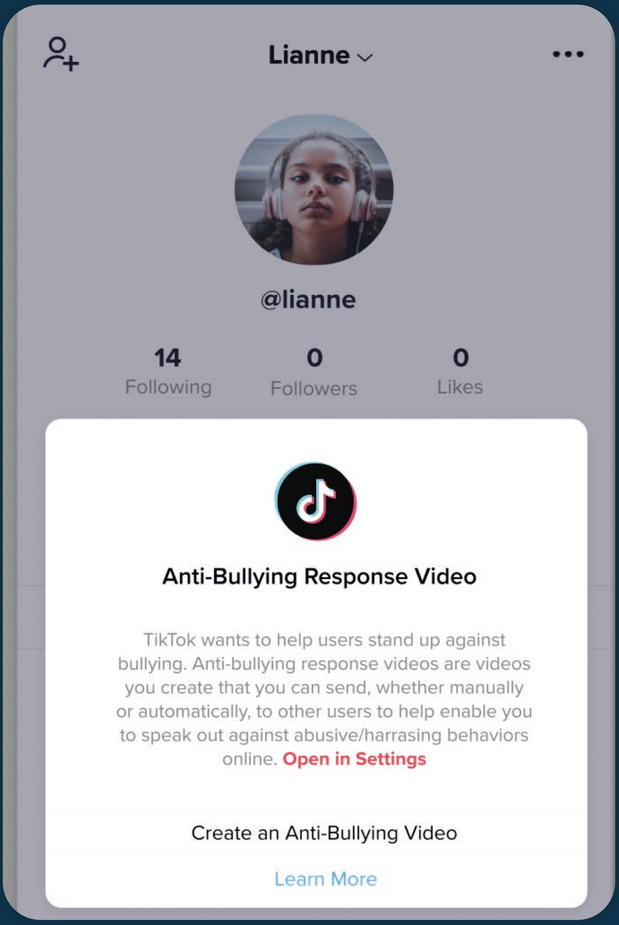
How would you categorize this message?

Figure sequence 5 (Demo 5): Anti-bullying video on TikTok



Tap to create a response to harassment and bullying

Tap Here for More Info

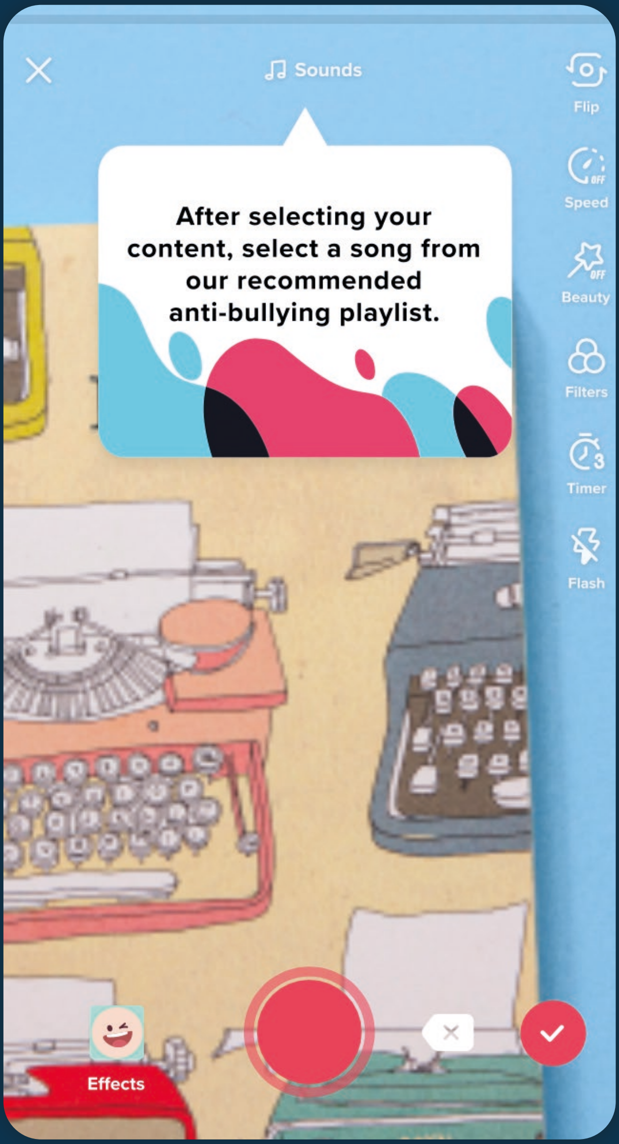


Anti-Bullying Response Video

TikTok wants to help users stand up against bullying. Anti-bullying response videos are videos you create that you can send, whether manually or automatically, to other users to help enable you to speak out against abusive/harrasing behaviors online. Open in Settings

Create an Anti-Bullying Video

Learn More



After selecting your content, select a song from our recommended anti-bullying playlist.

Method and data analyses

We rely 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 6 focus groups (from now on FGs: 4 groups with female participants conducted offline in one school in an urban area of Ireland, and 2 online FGs with males, with 6-10 children per group). See Tables 1 and 2 in the Appendix for the sample structure. All research was conducted in Ireland. Interview recruitment took place with the help of the youth organisation Foróige[20] as well as via Amarach research agency. The fieldwork was conducted from May to August 2021 and all except for the 4 school based FGs were conducted online due to lockdown conditions. All procedures received approval from the Dublin City University Research Ethics Committee (REC) as well as the Data Protection Unit (DPU). Parental/caregiver written consent as well as child written assent were sought from all participants following the provision of plain language statements (PLS) which explained that research was voluntary in nature and that they could give up at any time, as well as the principles of confidentiality and anonymity. The PLS stated these in a child friendly manner. Following the transcription and anonymisation procedures, three coders engaged in an iterative, thematic analysis of the data; they discussed the themes that emerged and refined broad themes into more nuanced ones, and discussed any disagreements as to how the content was coded (Boyatzis, 1998; Braun & Clarke, 2006). Deductive coding (following predefined codes) was performed first with all three coders searching for the research questions-driven codes; and open ended, inductive round of coding was performed thereafter, with coders adding codes that they thought emerged from the research, which were subsequently exchanged and discussed.

Limitations: We experienced significant recruitment difficulties and delays due to Covid-19 lockdown circumstances, and we were unable to specifically recruit children from non-white Irish ethnic backgrounds; while some children in our sample did come from minority ethnic backgrounds, we were not able to recruit based on this criterion nor did we consequently record this feature as a variable in our study. We were also unable to recruit any children who openly identified as non-binary in terms of their gender or as LGBTQI+ and non-governmental organizations in Ireland catering to this minority group were not able to assist with recruitment at the time of our fieldwork.

[20] Foróige. (n.d.). Foróige's philosophy. Retrieved from: https://www.foroige.ie/

Findings

Understanding cyberbullying

Participants had varied understanding as to what cyberbullying was. We deliberately did not wish to provide them with a definition in order to get a sense as to what they thought cyberbullying was and their perceptions of cyberbullying; and whether the incidents shown in demos were in fact cyberbullying from their perspective. We encouraged them to discuss cyberbullying by emphasising that there were no right or wrong answers. Some children knew that cyberbullying involved repeated harm and intentionality, but they nonetheless thought for example that the exclusion scenario (Demo 2) constituted cyberbullying even though the act of exclusion happened only once. Cyberbullying was frequently associated with something "deliberately mean" whereas harassment was sometimes seen as something that was annoying (not necessarily mean); they were unclear as to whether they perceived harassment or cyberbullying as more severe. Interestingly, there was no consistency in answers, some children provided contradictory explanations in a single interview. Nor could we find any patterns in their preferences as to whether they wanted to have one option to report all behaviours (as in "abuse, harassment or cyberbullying" – one label for all behaviours); or to be able to report abuse, harassment and cyberbullying separately – seeing them as distinct behaviours, and what value such distinction would add to them. Younger boys in FGs (13-14) had a clear preference for more options to label different behaviours. They did know that cyberbullying included repetition but it was not clear how that was different from harassment to them. It emerged from interviews with girls that having several options to label content would be helpful.

Girl: I think if we had them all and then could pick instead of picking one you could pick two

Interviewer: Oh, instead of picking one you could pick two, ok interesting. Because sometimes… why is that?

Girl: Because when you pick one, say 'bullying,' and then it could be in the mix between bullying and abusive. (Girl 1, 15, interview)

Yeah but it kind of – I don't really know they're both just not very good and I guess it depends on the person how they view the comments whether or not it's bullying or harassment. (Girl 2, 15, interview)

Well uh I think bullying is something that constantly happens, like they constantly keep doing it, and then uh harassment is uh kind of constantly sending someone a message like not necessarily a mean message but just something that'll make people kind of feel bad about what they posted and then I'd say abuse is a few mean messages coming through every once in a while. (Boy, 13, interview)

I think they're the same, I think – I don't know I think it's the same thing, I think like if someone is going out of their way and something bad about you I think it's bullying and like I don't – I think bullying is just very, it's bullying or it's not I don't think there [sic] like a difference, you get me […] yeah I think so yeah I think if you say any nasty thing it's bullying, harassment, it's all the same I think. (Girl 16, interview)

Uhh they're similar but they're not like the exact same. Well, like, not really but like they're all kind of similar because they're all mean in like their own way. (Boy 16, interview)

Overall, participants expressed mixed views as to their perceptions of effectiveness and desirability of the proposed AI-based interventions; as well as regarding their implications for privacy and freedom of expression (from now on: FoE). There were no notable variations in children's views in terms of their age and gender with some exceptions for specific interventions (such as the anti-bullying video), as discussed below. Furthermore, girls interviewed in school-based focus groups were more likely to criticise the effectiveness and desirability of any AI involvement when compared to boys interviewed in focus groups, and to all children interviewed in one-on-one interviews. This may have been affected by the dynamics of the focus groups; one factor that needs to be born in mind is that all one-on-one interviews were conducted with children online (on zoom) due to Covid-19 safety precautions, and the Ethics Committee required that a parent/guardian be present nearby during the online interviewing process, which may have affected children's answers in terms of social desirability to an extent as well (Miller et al., 2015).

Do children welcome AI-based cyberbullying interventions?

The majority of children in both interviews and focus groups expressed mixed views about whether they would welcome AI-based monitoring for the purpose of cyberbullying detection. Most of them said that they would welcome proactive AI-based scanning/crawling/monitoring or "AI working in the background" as we tried to explain the process to them, for the purpose of detecting cyberbullying. However, when asked about privacy concerns, they had second thoughts about the process. Overall, they would welcome AI-based support as long as they have the ability to opt in and out. Children were worried about the crawling of private/direct messages for the purposes of cyberbullying detection, especially as regards to the messaging services such as WhatsApp, which they saw as avenues for private communication. However, a smaller portion of children held the view that having AI-based monitoring on private messages is good because they thought that if one wanted to bully someone, they would do so via direct messaging. Participants were also concerned that the AI might get things wrong, and that joking and friendly banter might be detected as "bullying," therefore restricting legitimate speech (see Ging & O'Higgins Norman, 2016); and blowing things out of proportion, even creating unnecessary conflict. They also pointed out that one cannot rely solely on AI but that a human needed to be involved in the process of reviewing the content and providing help to the victim. AI was sometimes imagined as a "robot" especially by younger children, and as such, incapable of having the necessary level of sensitivity and empathy that a human would be expected to exhibit, and therefore unable to adequately respond to bullying.

> No not really, everything is alright [when asked if they would mind AI-based content monitoring]. I think the AI could help but I wouldn't rely on it solely to do that stuff. (Boy, 14, interview)
>
> No. I wouldn't mind, no. [when asked about AI in private messaging]. I think it's better, I think that's how it should be on direct messages, because if I was going to send a nasty message, I wouldn't comment it on someone's post, I would um send it directly. (Girl, 16, interview)
>
> **Girl 1:** No [it is not ok] ... That's an invasion of privacy [monitoring DMs]
>
> **Girl 2:** Well like, comments are public so yeah, that's fine. But messages, like unless they're reported..."
> (Girls FG, ages 16-17)

Comfortable and on-board [with AI monitoring]. (Girl, 12, interview)

I think you should be able to like allow them or just not allow the AI to do that. (Boy, 12, interview)

Girl 5: Also, what happens if it was just like, friends, joking about that. Because sometimes friends do that and be like, oh, you know...

Girl 2: Yeah like fat-shaming yeah.

Girl 1: Ah yeah and like slang and stuff, that's like, here, you look massive. And then there, it's like "you look fat". (Girls, FG, ages 16-17)

Girl 4: He [AI] could take [down] everything and anything at this rate. Like you comment something good and it could still report ya. Like you could literally post anything. And posts are getting taken down.

Girl 5: Like videos being removed for no reason and all. (Girls, FG, ages 13-14)

But I wouldn't want it [person doing the monitoring] to be a robot, I would want it to be a real person. Because robots do not have emotions. For example, if Susan from Germany is reviewing the content for bullying, and she might have been bullied previously in her life, so she will be able to understand that let's give that person a punishment. Whereas the robot on the comments like "You are loser", wouldn't think it is enough for account suspension or something. I feel like with humans, you can connect more personally, and people can feel pity on you. If a robot does it, it is just going to glance over it once and determine it within seconds that whether that was bullying or not. (Boy, 12, interview)

Well firstly I just want to say it is so weird to say that everything you text, someone is monitoring it, you think you're having these private chats but really you're not. it's actually scary, it's very scary and now when I go home I'm definitely going to think twice about what I even say in a private message, that's mad. (Boy, 16, interview)

Facial recognition: "creepy"? Perceptions of social media surveillance

Demo 2 leverages facial recognition to detect exclusion, and children expressed unease around the idea of using facial recognition, even if it is for the "greater good" of cyberbullying detection. Some children, however, did point out that they did not find it creepy because the technology was being deployed for a good reason – to detect cyberbullying. Most of them were not aware of facial recognition at all and that it was possible to detect their identity from their facial features. At the same time, several older children pointed out that there is no privacy online anyway, that they assume that everything they write and post, even if it is in private messaging is available to companies or governments anyway. They thought that their peers may not care about monitoring because they probably assume that everything is monitored anyway.

Girl 1: And how is Instagram able to detect that she wasn't in the photo? Like how does it do that?

Girl 2: That's creepy! (FG, girls, 16-17)

It's a bit weird how it can tell if you're tagged in a post… like how it knows your face. That's kind of like an invasion of privacy on its own. (Boy, FG 15-16)

Interviewer: You wouldn't let it scan your face for the sake of catching cyberbullying?

Boy 1: Like I wouldn't.

Boy 2: I wouldn't either. (Boys, FG 13-14).

I think it's creepy if it's scanning you for no reason [but] because it's actually trying to stop bullying, I don't think it's creepy. (Girl 16, interview)

Uhm well it [facial recognition] is kind of creepy to think about that it can do that, but in some cases it can be handy yeah it could kind of feel like an invasion of privacy but if you think about like the positive uses for this then it could kind of outweigh that feeling of an invasion of privacy (Girl 1, 15, interview)

Emm it doesn't really bother me that much I think. I think part of going on social media is knowing that a lot of it is monitored once it's put it there. (Girl 2, 15, interview)

I suppose that's kind of what it is and you forget that we're online because you're in this setting, it's done so well, that you feel like you're having this private conversation with someone but you're just in a chatroom with people. I'm on zoom with you and although you're recording it, I'm sure there is someone at the zoom headquarters making sure that something [meaning bad or bullying] is not happening here. (Boy, 16, interview)

Perceptions of effectiveness and desirability of the support contact/helper/ friend feature

The majority of children who took part in the interviews would overall welcome the option to add a support contact/helper/friend. While many said that they would use it, not all of them were sure that they would do so if given the opportunity; and they were also concerned that their peers might not use it. Girls in FGs and especially older girls (age 15-16, 16-17) had many concerns about the support contact option for a variety of reasons, and overall thought that this intervention was unnecessary, and that it could even make things worse in a cyberbullying incident. Older girls in FGs thought that such an option was more appropriate for younger children who could then add their parent/guardian, older sibling or friend as their support contact. Among the concerns children had were the following: 1. They may not feel comfortable admitting that they need a support contact/asking for help in a bullying situation. 2. They prefer to deal with cyberbullying when it happens to them on their own, which gives them a feeling of self-efficacy and prefer to just rely on tools such as untagging that allows them to deal with the situation on their own, rather than bringing other people into the incident 3. They thought that support contact might be overwhelmed with the requests for help and they suggested capping the number of people that one can be a support contact for 4. They were concerned that if a support contact is a friend of the victim, too many requests for help could annoy the support contact and damage the friendship. 5. They thought it was unfair to ask someone else to deal with one's own problems. There was also a sense that such support tools are more appropriate for more sensitive children that get easily upset and this sensitivity implied weakness which they did not seem to wish to identify with. Nonetheless, it is worth emphasising that these concerns were more prevalent in female FGs, especially older ones, whereas in male focus groups and one-on-one interviews, children appeared to be overall more welcoming of the support contact feature.

Interviewer: Yeah, that [support contact] would be something you might use?

Boy: yeah

Interviewer: And anyone else what would you think?

Boy 2: The thing is that is a good idea, it makes the person getting the comments not be alone. I suppose. (FG, Older boys, 15-16)

Think it's actually a really good idea cause like a lot of people can feel alone when they felt harassed or bullied online and to have someone there to see it and say hey it's alright and it's a really, really good idea (Boy, 13, interview)

Emm I think it's good I'm not 100% sure if teens would use it. I think teens often struggle with asking for help, but I think sure people maybe. Although TikTok does have an age like requirement I think a lot of younger kids would use TikTok and I think it would be very helpful for that age bracket. So, I think say 10 to even 13, 14 I think it would be lot more helpful (Girl 1, 15, interview)

Girl: I think it would be a wonderful idea.

Interviewer: Yeah, why?

Girl: Because some people might not want to talk themselves. They might want to fight for them. (Girl 2, 15, interview)

I think that's a good idea. You can get more support and help than just one person (Girl, 12, interview)

Interviewer: You could see yourself using that and maybe asking a friend to be a support contact?

Boy: mhm yeah

Interviewer: And do you think that the idea of a support contact would be good across different networks as well? Or just for TikTok?

Boy: Yeah I'd [say] Instagram cause there's a lot more comments on things like Instagram than on TikTok videos yeah. Yeah I think that was pretty good to have human input (Boy 14, interview)

Girl: Yeah [support contact could be a good idea] like if there was someone young on it [TikTok] and you wanted someone old to… [be a Helper]. (Girls FG age 13-14)

Girls in FGs and especially older ones were particularly vocal about the fact that involving AI and the support contact option in the exclusion scenario (Demo 2) was unnecessarily complicating the situation and making a big deal out of something that should have otherwise been easy – such as untagging oneself. They also thought that perhaps such measures would be appropriate in what they perceived were more severe cases such as racism or homophobic bullying, as opposed to a relational issue where someone called someone else "ugly," which they perceived as less serious. It was suggested that one should learn to deal with it on their own, that one mean comment does not mean much or that one should grow a bit of a thick skin if one is to be one social media. Furthermore, girls in older FGs and some children in interviews expressed concerns about messaging the perpetrator with a polite request to stop the abusive behaviour and/or an explanation that their behaviour is hurtful. Those with such concerns thought that if one wants to bully you, letting them know that they've managed to hurt you will only make them feel pleased that they've gotten to you; and moreover, that they'll likely go on with abuse.

Finally, they also thought that receiving a notification/alert that AI has detected abusive/potentially bullying content is unnecessary and that even if they are given the option not to see the content of the abusive message, they did not wish to know that this was happening. It is worth mentioning that not all children felt this way and that some children in interviews said they would want to be alerted when bullying happened.

Girl 1: Even like setting up the whole thing, whatever it is, yeah. I feel like it would be a lot to go through when, like, no matter what, people are gonna post. Even if its in their heads. So it doesn't really matter in the long term. Like, one comment on some posts. Like it's a good idea in theory, but like, I feel like even having to set up the whole thing I think a lot of people wouldn't bother to set it up… like I think it also takes case by case because like, if it was someone like actually like being racist or homophobic or something like that, like, that's a different thing… Its not with someone calling you ugly. You know what I mean? Like it would be upon that person.

Girl 2: I think like having the flagging is good, I don't know if I'd set up a Support…

Girl 1: A lot of people wouldn't go through the hassle, even though it's not that much of a hassle, people are lazy, It's a good idea in the big picture… But like, I feel like, it wouldn't be that like everyone would use it… You know what I mean?

Girl 3: I think it would be useful to certain people… Personally, I wouldn't use it. (FG, Girls, 15-16)

Girl 1: You're gonna have to… some of the comments at least… you're gonna have to be able to put up with it like… and just delete a comment. And not let everything get to you. You know what I mean? (FG, Girls 13-14)

Girl 2: Just untag her!

Girl 3: Why is she asking her friend? I just I don't think it's any of her [victim friend's] business, why is she like, asking her friend that? Like if you have a problem with someone just go say it to them, why are you bringing another person into it. […]

Girl 4: I thought that… that from the start, she had the option to just untag herself and like, she could have just left it there, like from the very start.

Girl 1: You already can do that.

Girl 4: Yeah that's what I'm saying, like it should be left there. Like, just untag yourself and go on about your day, like. (FG, Girls, 16-17)

I am not sure [about asking the perpetrator to stop] because, lot of the time I think it could work in some aspects that if the comment is mean in a way they don't mean it. So, I think certain comments it's kind of not they're not supposed to be rude but they can come across as rude so I think in that way it might help; but in a lot of instances I feel like asking him [to] stop isn't gonna help. (Girl 1, 15, interview)

Girl: Yeah, I would do this [ask the perpetrator to stop/take something down on behalf of a friend – as support contact].

Interviewer: If you were prompted by artificial intelligence?

Girl: Yeah, I would definitely do it

Interviewer: You wouldn't feel uncomfortable or sort of kind of reaching out to these people?

Girl: No, I wouldn't feel uncomfortable

Girl: If I think your friends got her involved it wouldn't be as bad… but because TikTok got her involved. (Girl 2, 15, interview)

AI-prompted bystander involvement

The option for AI-triggered bystander involvement was not met with much enthusiasm. Once AI detected cyberbullying, the person who saw the post or commented something positive or neutral on the post which received abuse would be alerted of the fact that negative comments had been posted, and their help would be solicited. Some children in interviews thought that it would be a good idea for the victim to receive more help, but they also pointed out that it is platform dependent. For example, on TikTok in particular and on Instagram as well, many posts could be seen by or receive neutral or positive comments from complete strangers who have no interest in getting involved. If these people were to suddenly receive alerts that random strangers were being bullied, they might even be annoyed with all the notifications, children surmised. Furthermore, some of these bystanders might be inclined to support the perpetrator for fun or for some other reason, and therefore children thought that alerting random strangers of bullying is a risk. Some of them also pointed out that they would not wish to involve so many people and that they would rather deal with these issues on their own. There was also a sense that people say mean things without thinking about it or thinking of meanness as a joke "slagging" (see Ging & O'Higgins Norman, 2016), then they tag their friends so that they can laugh too; if all of those people [without further scrutiny as to who they were] were to be notified, the incident could be blown out of proportion, even to the detriment of the victim.

Girl 1: She's [bystander] trying to be help and be nice but it's nothing to do with her… People are there calling her [victim] fat and she's [bystander] giving her positive energy.

Girl 2: Like if everyone who commented on that and then got a notification…

Girl 3: A lot of people who comment take their friends or whatever… But if brought back into it, they could make it worse…

Girl 4: Okay so if your page is public you can't differentiate how many people will see it, how many people will comment… It could be someone you don't know who comments something nice about you. And getting them involved is a bit… I don't get it… (FG, Girls, 15-16)

Girl 1: It's a bit unnecessary…

Girl 2: I don't really get why TikTok would put someone who doesn't even know the other person like that's not their business.

Girl 3: Yeah if TikTok did that… and see what happens… it's a bit stupid.

Girl 4: Its nothing to do with her really.

Girl 1: Maybe she might have commented before it or maybe after… But she mightn't have seen the comment dya get me? She might not even know that person! (FG, Girls, 13-14)

Uhm well maybe (she might use bystander feature), cause you see that some of the people, like if let's say you're contacted as a bystander on this you might not have actually seen it, you could have commented or the other person did, but you know uh it like if you – I don't really know but I'd say it's decent you know and that person could help again, you know, prevent the negative comments. (Girl, 15, interview)

Maybe not the bystander feature, that's not the best idea but the support contact that's actually really cool. (Boy, 15, interview)

I think it's a really good idea [bystander] because say if the person, like the support contact, didn't know what to do they could get another person in to try and help them make a decision of what they should do. (Boy, 13, interview)

I don't think much [sic] people would like to get involved because I don't know, they wouldn't really know the person, so they wouldn't take it personally, whereas if they were like best friends with Sally [victim] then they probably would say yeah, but people wouldn't know each other. (Girl, 12, interview)

School involvement

Children expressed ambivalence about reporting cyberbullying on social media to their school and about school involvement in general. While many children in interviews said it would be a good idea to have the option to report to school in theory, a number of them brought up possible caveats that would prevent them from doing so; or why they thought their peers might not wish to do it, such as: They thought that there was little that the school could do in a situation where the perpetrator did not attend that particular school; or if the incident did not happen on school premises or was somehow school-related (some thought that the school was not responsible in such circumstances eg if the incident happened outside school hours). Some children were sceptical that reporting to teachers and school staff was effective even if all participants in a cyberbullying incident attended the same school (some even thought the school involvement makes things worse in bullying incidents). Fewer children said they would rather report to their teachers in person; or that they do not like the idea of school having anything to do with their Instagram or social media presence. For some children, it was important to be able to report to school anonymously. It is worth pointing out that according to the latest nationally representative data from Ireland, 82% of parents/caregivers said they would prefer to receive online safety information or advice from their child's school and 60% said they did so already (National Advisory Council for Online Safety, 2021). The fact that many children thought schools were not necessarily responsible for incidents that happened online, outside of school hours or premises, reveals that they are not aware of the policy framework in Ireland which provides the school with a remit to become involved in online incidents and out of school experiences. Schools do have a remit in relation to out of school bullying and cyberbullying in so far as these may impact on a child's right to access and participate in their education (Action Plan on Bullying, 2013[21]; Anti-Bullying Procedures for Primary and Post Primary Schools, 2013[22]; Children First Act, 2015[23]; Child Protection Procedures for Primary and Post primary schools, 2017[24]).

[21] Minister for Education and Skills, IE. (2013). Action plan on bullying. Retrieved from: https://assets.gov.ie/24758/0966ef74d92c4af3b50d64d286ce67d0.pdf

[22] Circular 045/2013. Anti-bullying Procedures for Primary and Post Primary Schools. Retrieved from: https://circulars.gov.ie/pdf/circular/education/2013/45.pdf

[23] TUSLA, Child and Family Agency. (n.d.). Children First Guidance and Legislation. Retrieved from: https://www.tusla.ie/children-first/children-first-guidance-and-legislation/

[24] Child Protection Procedures for Primary and Post-Primary Schools, IE. (2017). Retrieved from: https://www.pdst.ie/sites/default/files/Child%20Protection%20Procedures%202017.pdf

Girl 1: The school will very unlikely respond to that as well…

Girl 2: There's not really much they [school] can do… They can say "stop fighting"

Girl 3: Yeah but then they'll just go "say sorry" and leave it at that…

Girl 4: Yeah and that wouldn't really solve it…

Girl 1: But the school be like… it DID happen in school… "we can't do anything about it"… "bring it to your parents".

Girl 5: And that's because they have to, not because they want to! (FG, Girls 13-14)

Girl 1: If it's two people in the school yeah no it shouldn't matter… But if it's a different school…

Girl 2: They can't really do anything!

Girl 1: But even if it's in school grounds then it's okay.

Interviewer: But if it was a Saturday or somewhere like…

Girl 1: Somewhere like McDonald's yeah…

Girl 2: They can't do anything.

Girl 3: I think if it's two students the school can [do something]

Girl 4: Yeah even if it's a picture of two students they can do something… they [the school] have some control of it. (FG, Girls, 15-16)

Uhm well it can be handy especially if it happened at school uhm or if you it depends if it's involved and has something to do with the school itself or if you don't want to – if you don't want your parents to deal with it for you uhm cause maybe some people might prefer to go to their school and when this stuff happens to them. (Girl, interview, 15)

I really like it, I really like that. My new school – they would deal with it properly and they would look at the report and they were very helpful, but you can't guarantee that every student goes to a good school. (Boy 16, interview)

Interviewer: What do you think about that? To report to the school anonymously

Boy 1: Yeah it's good, it's useful.

Interviewer: Yeah, I'm hearing it's useful. What do others think?

Boy 2: I think it's' good.

Boy 3: I think it's good as well. (FG Boys 13-14)

Support score/unlocking features: nice but unnecessary?

Rewarding upstanding (support for those who are experiencing cyberbullying victimisation) with support score points or with more platform features (colours, emojis) was generally seen as desirable but many children had concerns as to whether this would work on social media platforms other than Trill, which was shown in the demo. Some said such rewards could incentivise them and their peers to help. Nonetheless, some children thought that it is a bit strange to be rewarded for helping – something that goes without saying that one should do for their friends. There was also a perception that stepping in was something that one was expected to do for their friends even without being rewarded for it, but that a reward was not enough of an incentive for helping a complete stranger unless the bullying case was very severe.

Because, rewarding people encourages people to improve and punishing people it's like letting them know, what you did isn't going to slip by us. (Boy, 12 interview)

Yeah, cause if I was like oh I might not bother [to help], I don't want to get involved, but I could get some points for this, even though it's bad to say that but like it's true. (Girl, 16, interview)

Yeah, it'd make people more involved and stuff and the fact that they can get bonuses and extra points or whatever it's called on different apps. (Boy, 16, interview)

Girl 1: You might only want to get points to get points not to actually help. Someone might report something so you can be rewarded.

Girl 2: Some people will just report [any user] to get the colours... (FG, Girls 15-16)

Girl 1: 10 points for being nice and being a friend?

Girl 2: And it should just be that anyway.

Girl 3: Exactly! (FG Girls, 13-14)

Boy 1: It's a good idea [support score and unlocking features]. I think it will get people to kind of like, be more honest, like helping...

Interviewer: What do you think?

Boy 2: Yeah, I think it's very decent.

Interviewer: Yeah? Would you see yourselves, like proactively getting engaged if you are going to be rewarded for it?

Boy 2: I would yeah...

Boy 3: Depends on what they were rewarded for...

Interviewer: Let's say its a comment, I guess, where we can see that it's purely targeted at like your mate or someone like that for like being themselves...

Boy 3: It was my friend. Yeah.

Interviewer: And even if it was another user, would you do it?

Boy 3: I mean, maybe depends how bad the message is really. Like, if it's like sending like things like, oh, go kill yourself... or rape... or something like that. Yeah. Like I'd step in and then, but if was just insulting [then no]. (FG, Boys 15-16)

Anti-bullying video

Anti-bullying video was met with mixed feelings since older children in particular thought that it could make things worse for those who were bullied, even that it was "cringey." Most older participants in FGs and interviews thought that it was more appropriate for younger children. Fewer participants liked the feature saying that they would actually use it, that it was cool and that it would be effective. Most of them said they preferred to decide on a case-by-case basis, whether the video should be sent to the perpetrator if they experienced cyberbullying; rather than having the video sent to the perpetrator automatically once AI detected cyberbullying. There was a sense among older girls that if the feature were to become "cool" among popular or even famous people, then it could become appealing to a wider population; and that it could be handy to have it sent automatically especially by those who do not have the time to deal with many mean comments.

Girl 1: That's a bit much!

Girl 2: What's a bullying video gonna do? Like ya can just "flick off it."

Girl 3: And if you're gonna say to a bully "you're hurting me" well like "yeah that's the point!" (FG, Girls 13-14)

Boy: I think it is a really good idea so 'cause like it could well if it was like automatic and it was sent to people I think they would really make people think about what they've said or done and probably make them change their ways or apologize;

Interviewer: Would you like to have this option of the video on TikTok?

Boy: I'd say so yeah

Interviewer: Would you use it?

Boy: Probably (Boy,13 interview)

No, I think if someone's making fun of you or whatever and you send them this video, I think they might just think it's cringe and then it'll backfire, I don't know. (Girl, 16 interview)

I think people would make, I think people make fun of it. I think was to send a hurtful message and they got that response [sic]. I think it could almost feel them to say more hurtful things (Girl, 15, interview)

Girl 1: Certain people will do it [use automated anti-bullying video option] more because they don't have time to deal with a lot of abuse...

Interviewer: Younger or older or... appeal to anyone?

Girl 1: I think it's there's plenty of people that I really, I don't know how to explain it because they don't have time to like, reply to everyone every single hate comment... So if it's automatic...

Girl 2: And then as soon as they see famous people using it, they'll all start using it. (FG, Girls 15-16)

They might think you're weird sending a video or something. (Girl, interview, 12)

Yeah, I would say so, it might be done, but I just don't know how seriously it would be taken. I'd say that a bully has in their head "haha it's so funny I'm bullying, I'm saying what I want to this person" you know, who are they, they're a weirdo, they're whatever. And if they send that to them they'll go "haha who do they think they are, they're so weird" and it will just continue. I don't know how realistic it is though. (Boy 16 interview)

Yeah, I think they would be meaner if they knew that you were going to do that [send anti-bullying video] and they just found it funny that you would do that so they are mean about that then too. (Boy, 12, interview)

Reflective messages

We have only managed to show reflective messages in two interviews as this was an optional demo and it was only shown if there was time left after all the other core and optional demos were inquired about. Both participants agreed that reflective messages could be useful but they were not entirely sure if such a measure would really prevent someone from engaging in CB if they really wanted to harm someone. As for the implications for FoE, a girl (15) thought that even though the reflective message was a prompt to a prospective post that did not contain bullying but rather an unflattering opinion, such measure was still appropriate because in her view, opinions that are potentially hurtful should not be stated at all.

Boy: I think this would help them to kind of realize what they've done wrong and correct their mistake.

Interviewer: And do you think if someone wanted to bully someone would they really stop because of this message?

Boy: I wouldn't say they would if I'm being honest. (Boy, 13, interview)

Girl: Uh I think that's good because it could get people to think before they end up posting it and posting something hurtful um because they might just do it but that'll kind of you know, make them think twice about it and you know try to get them to realize what they're posting, what they're commenting is hurtful.

Interviewer: And even if it's not like this is sort of it's just an opinion, right? It's not really bullying, or maybe you would disagree?

Girl: Well it's more of an unwanted opinion because if what you're saying doesn't benefit or if it's not nice you know it's like if you've got nothing nice to say don't say it at all, um so if it's if you don't think that it will help that person or be beneficial in a positive way to the person that posted that picture or video then you shouldn't really be saying anything. (Girl, 15, interview)

Punishing perpetrators with less engagement and freedom of expression (FoE)

As a reminder, we solicited children's views as to punishing the perpetrator of cyberbullying[25] by having all their subsequent content (regardless of its nature – it could be positive content not necessarily abusive) under prioritised by the platform algorithm for a month and therefore less visible to their followers, friends and other users (similar to shadow banning except that the user is notified of the fact that their posts are receiving less engagement). We wondered if children would find that perilous from the perspective of FoE and whether they thought it was an effective of a punishment. Children did not find it particularly concerning that the perpetrator should be punished in this manner from the perspective of FoE. As long as the perpetrator had the right to appeal the decision, and as long as the punishment was time-limited, or triggered only after repeated perpetration, they would welcome such an intervention on social media platforms. A few of them stressed, however, that removing the perpetrator from the platform or having their posts banned altogether would be more effective of a deterrent.

Boy: I'd say maybe for like a few days or like a week or something would be like uh a good um punishment but not for like say a month or all posts after that should be limited.

Interviewer: Why so?

Boy: Ugh because I think say if the person that did do the harassment did realize oh I shouldn't do this and then they apologize to the person they were harassing (Boy, 13, interview)

[25] After the post was detected by AI, reported to the platform and confirmed as violating platform policy.

Girl: I mean after one post he shouldn't get punished that much like if he did it three times

Interviewer: Ok like a three time strike kind of thing?

Girl: Yeah

Interviewer: Ok, yeah. But after three times you think it would be a good idea?

Girl: Yeah (Girl, 15, interview)

It's quite good. It probably would [work]. Yeah. Because they [perpetrator] want as many likes and as many followers and stuff as possible. (Girl, 12, interview)

I feel like for some people it's always about followers, views, likes. I mean, he could also just not care about it. Some people are just obsessed with views, followers and likes, and some people couldn't care less. (Boy 1, 12, interview)

There should be another form of punishment in place. Like account suspension or something, or something like he's not allowed to post anything. (Boy 2, 12, interview)

Girl 1: Freedom of speech isn't freedom of consequence. [...]

Girl 1: Yeah. Yeah, like, there's freedom of speech. But you have to respect others as well. You can't just use that as an excuse to get away with everything.

Girl 2: Because like, you can't just go around saying, oh, freedom of speech. Yeah. That's just not on! (FG, girls 13-14)

Girl 1: Because like, if his account is like, it could be a burner account as well, like a burner account you can use to make fun of people. So, if it's just going to have less engagement, it doesn't really affect him. [...]

Girl 1: Like the less... the less engagement thing, if the person was popular, then obviously it might be annoying to them. But for it to be like the default punishment, especially if it's someone with a small account, that doesn't affect them in the slightest.

Girl 2: I don't really like the punishment, because like, if you posted a mean post about someone, I think that post should be dealt with, and if it's a case where it's like something you've done in the past, like you're like a repeat offender, then I think the lesson engagement thing is a good idea. But if it's your first time, like just making fun of someone, I think there should be like a different type of punishment. Like I don't think the less engagement is... Because what if you don't care about engagement, and you just want that one person you're bullying to see like, you don't care if anyone else sees. [it] (FG girls 16-17)

Interviewer: So in this situation, do you think shadow banning would be effective?

Boy 1: Kind of yeah. I mean it's less, not none. So he could send it [mean message] to someone else.

Boy 2: I mean you can just make a new account and do the same thing.

Boy 1: Like it's Instagram, it takes a maybe a second to set up a new account. You get a burner email that doesn't link back to you... picture that doesn't link back to you. Or just no picture.

Boy 2: You can use one of those random email generators. (FG boys 15-16)

We also proposed an intervention which would allow the support contact and the victim to prevent the sharing of the content that was labelled by AI as cyberbullying on other social media platforms until the content was reported and reviewed by the platform itself to be taken down if determined to be against the platform policy. Such a measure is intended to stop the spreading of hurtful content. Children who were asked about this option generally thought it would be a useful option to have and did not raise FoE concerns on their own. Nonetheless, many of them pointed out that it would not necessarily be effective since anyone can take a screenshot and find other ways to share the same content.

Concluding summary

In this report, we detailed the key findings of the project which solicited children's ideas and suggestions as to the design of AI-based cyberbullying interventions on popular social media platforms. We were especially interested in children's views as to how proactive regulation of abusive behaviours such as cyberbullying affected their rights to protection (safety) and what kind of interventions they would consider to be effective in reducing cyberbullying; how these interventions affected their rights to privacy, freedom of expression and access to information. Unlike with reactive moderation, where a child first reports content or an account to the platform before moderation takes place, proactive moderation refers to platforms deploying AI to detect and take action against abusive behaviours before they are reported by users (Milosevic et al., 2022).

Following the UN Committee on the Rights of the Child's adoption of the General Comment No. 25, children's rights, as stipulated in the UNCRC, apply in the digital world (Livingstone, 2021; Third et al., 2021). This means that in addition to the right to protection (safety), privacy and freedom of expression, they also have the right to be heard on matters that concern them (Article 12). While states and not technology companies are the primary duty bearers of the UNCRC implementation (see eg, Benesch, 2020), the passage of the General Comment No. 25 nonetheless underscores the calls long made by scholars: that all stakeholders whose activity has an impact on children's lives, including technology industry, need to take responsibility to improve children's rights in digital environments; and especially to take into account children's views when developing polices and mechanisms that impact children (Lievens et al., 2018; Staksrud, 2016).

Embedding children's views on matters that concern them with respect to technology design will become ever more important with the implementation of national laws that regulate online safety such as the OSMR in Ireland and Online Safety Bill in the UK; as well as the Digital Services Act at the EU level. With respect to privacy and the General Data Protection Regulation (GDPR)[26] implementation, the Irish Data Protection Commission's Fundamentals for a Child-Oriented Approach to Data Processing already stipulates, among other clauses, that children should have their say as regards to data processing by online services.[27] In our study, therefore, we solicit children's views on AI-based enforcement as a step towards ensuring that child best interests are a primary consideration as regards to interventions that have a clear impact on them. The authors of this report have long been emphasising the

[26] GDPR. EU. (n.d.) Complete Guide to GDPR compliance. Retrieved from: https://gdpr.eu/

[27] Data Protection Commission, Ireland. (2021, December). Fundamentals for a Child-Oriented Approach to Data Processing. Retrieved from: https://www.dataprotection.ie/sites/default/files/uploads/2021-12/Fundamentals%20for%20a%20Child-Oriented%20Approach%20to%20Data%20Processing_FINAL_EN.pdf

need for technology companies to consult children's views during the safety policy design process, and moreover to make the information about how this is done and the results of this process open to public scrutiny (Milosevic, 2018).

Our results, based on qualitative research with 59 adolescents aged 12 to 17 from Ireland, suggest that children would generally welcome AI-based interventions provided that they are given the option to opt in and out. Children, however, brought up a number of privacy concerns especially as regards to the use of facial recognition and DM/private message monitoring for the purpose of cyberbullying intervention.

While most of them would welcome the option to have a support contact/helper/friend whose help could be solicited when cyberbullying is detected by AI, children brought up a number of concerns about the effectiveness of such assistance and willingness to use it: from preferring to deal with cyberbullying on their own; unwillingness to tell others that they experienced cyberbullying and to bring them into the incident; to the fear of burdening their friends with their own problems. Involving the support contact to ask the perpetrator to stop was considered to be particularly problematic, especially by older girls (15-17) in focus groups, as they did not think someone else should be responsible for solving their own problems. Some pointed out they would be reluctant to admit they have a support contact, as this implied weakness or lack of self-reliance and was perceived to be appropriate for smaller children.

Involving bystanders into becoming upstanders and studying the conditions under which they are most likely to become involved in helping children who are experiencing victimisation, is a widely researched issue (Bastiaensens et al., 2014; DiFranzo et al., 2018; Macaulay et al., 2022; Williford et al., 2013). In our study, children expressed reluctance to bring in random bystanders into the incident, emphasising that such involvement was platform and context dependent. They preferred to address the problem with their support contact or on their own and even said that bystanders (if they are strangers) could make things worse.

While many children were reluctant to bring others in, they seemed to think that if a bystander was someone they knew it could be a welcome idea, depending on the context of a particular incident.

Reporting incidents to school via an official Instagram account handled by the school counsellor or another professional, was also met with ambivalence; some children thought it would be helpful to have it in place but brought up a number of reasons as to why they would not wish to have the school involved. Some thought that there was little the school could do in any event, and especially if the perpetrator did not go to that school; or that the school was not responsible for what happened online outside school hours and premises.

Custom tailored anti-bullying videos which could be sent in response to abusive behaviour when detected by AI, were met with mixed feelings, seen to be more appropriate for younger children; and many children thought telling someone who is bullying you to be kind could backfire; just like telling them that their behaviour is hurtful could be counterproductive (children surmised that some perpetrators could think: "well, that is the point, I want to hurt you."). A number of cyberbullying interventions designed by adults, researchers and advocacy organisations, many of which are featured every year on the Safer Internet Day,[28] include messages such as "Be Kind!" and "Don't Bully". Feedback we received from children shows how these messages might fail to resonate with youth culture and how we need to ensure that cyberbullying prevention and intervention is meaningful and context-sensitive in order for it to be effective in reducing the problem (Jones et al., 2014; Finkelhor et al., 2021).

While respondents did not seem too concerned about FoE, they nonetheless emphasised the importance of effective appeals mechanisms when AI-based takedown decisions or activity restrictions are being made (such as the perpetrator's content being algorithmically underprioritized, similar to shadow banning).

Some pointed out such restrictions should be time-limited or triggered only after repeated violations and reconsidered after a while. While they did not think that less engagement should be the only punishment available, (some thought banning or content take-down was more appropriate), they would welcome this feature as long as it is transparent, and appropriate appeals mechanisms are provided. Similarly, they thought that giving the victim and support contact the option to restrict the sharing of AI-detected cyberbullying content to other platforms (such as posts or stories eg, from Instagram to Snapchat etc.) could be a welcome feature; however they thought it would not be necessarily effective given that one can screenshot and copy content in many ways. Finally, they were overall worried that AI would wrongly detect joking or slagging as cyberbullying, which would negatively impact FoE as well as their friendships.

Policy recommendations

Technology companies

- ▶ There are ample ways in which companies could conceptualise and design evidence-based proactive (AI) interventions, relying on technological advancements that are already available to them. We have suggested some of these interventions in this report, along with children's views on their desirability, effectiveness and possible alternatives.
    - ▼ Consider adding the support contact/helper/friend option on the sign up, especially for younger children, limiting the number of people anyone can be a support contact for; making sure that signing up is anonymous and need not be disclosed to anyone (other than the support contact, of course).
    - ▼ Bystander feature needs further research but holds potential pending further evidence; it would need to be carefully implemented within each platform's specific context.
    - ▼ Consider further research into school involvement either as trusted flaggers or as avenues to report cyberbullying incidents.

[28] Safer Internet Day. (n.d.) Together for a Better Internet. Retrieved from: https://www.saferinternetday.org/

  - Consider any future decisions as to applying facial recognition[29] and private message screening for safety purposes in the context of children's concerns about their privacy.
  - Consider further researching custom-made anti-bullying video as a possible optional feature for younger children; especially understanding how such a feature could be made desirable and normative from the perspective of young people.
  - Ensure that banning and restricting is done in a transparent manner i.e. that users know that they have been banned or restricted and why this happened[30] (decisions that involve under prioritising of perpetrator's content by the algorithm; limiting the sharing of content to other platforms); and that they have an effective appeals mechanism; consider further research into such interventions as supplementary to account banning and content take down.
  - Many children (especially older girls) appear to like the idea of handling cyberbullying on their own, quietly ("just untagging"), which is what platforms offer at the moment (empower the user by giving them as many options as possible to shield themselves from such content by muting, restricting, which does not involve platform intervention).
  - However, **it is also important** to conduct further research into understanding to what extent such preference is the result of normative conditioning (fear that one will be perceived as sensitive or weak if one asks for help); and ensure that the perceived demand for self-reliance is not suppressing children's authentic need and ability to ask for help.
  - Further testing of support scores and unlocking platform-specific features in the context of various social media platforms might be beneficial.
- With growing expansion into virtual reality, engaging in co-design[31] will become increasingly important to ensuring interventions that are effective and meaningful to children, and that honour the entire spectrum of their rights as well as the balance of these rights (such as protection and participation) as per UNCRC.
- It is important to make any such consultations and the results of them transparent to the broader research and academic community so that their effectiveness can be scrutinised by experts in the field.

[29] Exclusion-based bullying which was said to be frequent on Instagram: According to information presented at Meta/Facebook Global Safety Summit, 2019: https://about.fb.com/news/2019/05/2019-global-safety-well-being-summit/

[30] Companies are concerned that by revealing the exact details of their policies and their moderation decisions, they might inadvertently provide guidelines for those who wish to violate the policies as to how to get around those (see Milosevic, 2018). We do not think that by revealing to the user that a piece of their content or an action violated the company policy would necessarily lead to such an outcome. It is important to exhibit transparency in the context of restrictive decisions, and children have voiced such concerns as well.

[31] Meta. (2022, March 16). Introducing Family Centre and Parental Supervision Tools on Instagram and in VR. Retrieved from: https://about.fb.com/news/2022/03/parental-supervision-tools-instagram-vr/

- While content take-down and other moderation decisions involve assessing content against Terms of Service, Community Standards/Guidelines and other policies, soliciting children's views on safety, privacy and FoE implications can facilitate the design of interventions that make such decision making easier, and enable innovation.

Policy makers

- Policy makers have the responsibility to ensure the implementation of the UNCRC in digital environments and that children's right to be heard on matters that concern them applies to AI-based policy enforcement by technology companies.
- With the adoption of the OSMR in Ireland and DSA at the broader EU level, policy makers **have the ability to embed child consultations into codes of conduct**, which will be an integral part of this regulation; **as well as to involve children into the processes of auditing of companies' policy enforcement mechanisms** by soliciting children's views during the evaluation process.
- Such procedures can contribute to ensuring a balance of children's rights, decision-making that prioritises best interests of the child and improved effectiveness of policies and enforcement mechanisms from the perspective of children.
- Ensure that all school staff have clarity as to their duty of care towards children who are cyberbullied, regardless of the location of the perpetrator or the target. Schools do have a responsibility (remit) in relation to out of school bullying and cyberbullying in so far as these may impact on a child's right to access and participate in their education, following policy documents cited in this report.

References

Anderle, M. (2016, March 15). Making a more Empathetic Facebook. *The Atlantic*. Retrieved from: https://www.theatlantic.com/technology/archive/2016/03/facebooks-anti-bullying-efforts/473871/

Ashktorab, Z., & Vitak, J. (2016, May). Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 3895-3905).

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology*, *23*(1), 37-51.

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.

Bastiaensens, S., Pabian, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2016). From normative influence to social pressure: How relevant others affect whether bystanders join in cyberbullying. *Social Development*, *25*(1), 193-211.

Bauman, S., & Yoon, J. (2014). This issue: Theories of bullying and cyberbullying. *Theory Into Practice*, *53*(4), 253-256.

Benesch, S. (2020). But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies. *JREG Bulletin*, *38*, 86.

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.

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).

DeSmet, A., Veldeman, C., Poels, K., Bastiaensens, S., Van Cleemput, K., Vandebosch, H., & De Bourdeaudhuij, I. (2014). Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents. *Cyberpsychology, Behavior, and Social Networking*, *17*(4), 207-215.

Douek, E. (2022). Second Wave Content Moderation Institutional Design: From Rights To Regulatory Thinking. *Available at SSRN 4005326*.

Espelage, D. L., Rao, M. A., & Craven, R. G. (2012). Theories of cyberbullying. In S. Bauman, D. Cross and J. Walker (Eds.). *Principles of cyberbullying research: Definitions, measures, and methodology*, 49-67.

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.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Ging, D., & O'Higgins Norman, J. (2016). Cyberbullying, conflict management or just messing? Teenage girls' understandings and experiences of gender, friendship, and conflict on Facebook in an Irish second-level school. *Feminist Media Studies*, *16*(5), 805-821.

Görzig, A., & Macháčková, H. (2015). Cyberbullying from a socio-ecological perspective: a contemporary synthesis of findings from EU Kids Online. Retrieved from: https://www.researchgate.net/publication/281554815_Cyberbullying_from_a_socioecological_perspective_A_contemporary_synthesis_of_findings_from_EU_Kids_Online

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.

Heldt, A. & Dreyer, S. (2021). Competent third parties and content moderation on platforms. Journal of Information Policy, 11, 265-300. https://doi.org/10.5325/jinfopoli.11.2021.0266

Hinduja, S., & Patchin, J. W. (2013). Social influences on cyberbullying behaviors among middle and high school students. *Journal of youth and adolescence*, *42*(5), 711-722.

Hinduja, S., & Patchin, J. W. (2015). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin press.

Jones, L. M., Mitchell, K. J., & Walsh, W. A. (2014). A content analysis of youth internet safety programs: Are effective prevention strategies being used? Retrieved from: https://scholars.unh.edu/ccrc/41/

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, *140*(4), 1073.

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?*. Prentice Hall.

Lieberman, H., Dinakar, K., & Jones, B. (2011). Let's gang up on cyberbullying. *Computer*, *44*(9), 93-96.

Lievens, E., Livingstone, S., McLaughlin, S., O'Neill, B., & Verdoodt, V. (2018). Children's rights and digital technologies. *International human rights of children*, 1-27.

Livingstone, S., Carr, J., & Byrne, J. (2016). One in three: Internet governance and children's rights.

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.

Livingstone, S., Stoilova, M., & Nandagiri, R. (2019). Children's data and privacy online: growing up in a digital age: an evidence review. Retrieved from: http://eprints.lse.ac.uk/101283/1/Livingstone_childrens_data_and_privacy_online_evidence_review_published.pdf

Livingstone, S., Stoilova, M., Nandagiri, R., Milosevic, T., Zdrodowska, A., Mascheroni, G., ... & Wartella, E. A. (2020). The datafication of childhood: Examining children's and parents' data practices, children's right to privacy and parents' dilemmas. AoIR Selected Papers of Internet Research.

Livingstone, S. (2021, February 4). Children's Rights Apply in the Digital World! *LSE blogs*. Retrieved from: https://blogs.lse.ac.uk/medialse/2021/02/04/childrens-rights-apply-in-the-digital-world/

Lobe, B., Velicu, A., Staksrud, E., Chaudron, S., & Di Gioia, R. (2021). How children (10-18) experienced online risks during the Covid-19 lockdown-Spring 2020. *Key findings from surveying families in 11 European countries*. The Joint Research Centre (JRC) of the European Commission. Retrieved from: https://publications.jrc.ec.europa.eu/repository/handle/JRC124034

Macaulay, P. J., Betts, L. R., Stiller, J., & Kellezi, B. (2022). Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. *Computers in Human Behavior*, 107238.

Macháčková, H., & Pfetsch, J. (2016). Bystanders' responses to offline bullying and cyberbullying: The role of empathy and normative beliefs about aggression. *Scandinavian journal of psychology*, *57*(2), 169-176.

Marwick, A., & Boyd, D. (2014). 'It's just drama': Teen perspectives on conflict and aggression in a networked era. *Journal of youth studies*, *17*(9), 1187-1204.

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

Miller, P. H., Baxter, S. D., Royer, J. A., Hitchcock, D. B., Smith, A. F., Collins, K. L., ... & Finney, C. J. (2015). Children's social desirability: Effects of test assessment mode. *Personality and individual differences*, *83*, 85-90.

Mishna, F., Saini, M., & Solomon, S. (2009). Ongoing and online: Children and youth's

perceptions of cyber bullying. *Children and Youth Services Review*, *31*(12), 1222-1228.

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., Van Royen, K., & Davis, B. (2022). Artificial intelligence to address cyberbullying, harassment and abuse: new directions in the midst of complexity. *International journal of bullying prevention*, 1-5.

Milosevic, T., Verma, K., Davis, B., Laffan, D., Walshe, R., O'Higgins Norman, J. (2021, September). Developing AI-based Interventions on Online Platforms: Standardising Children's Rights. 11th International Conference on Standardisation and Innovation in Information Technology (SIIT).

Milosevic, T. (2016). Social media companies' cyberbullying policies. *International Journal of Communication*, *10*, 22.

Milosevic, T. (2018). *Protecting children online?: Cyberbullying policies of social media companies*. The MIT Press.

Montgomery, K. C., Chester, J., & Milosevic, T. (2017). Children's privacy in the big data era: Research opportunities. *Pediatrics*, *140*(Supplement_2), S117-S121.

O'Higgins Norman, J. (2020). Tackling bullying from the inside out: Shifting paradigms in bullying research and interventions. *International journal of bullying prevention*, *2*(3), 161-169.

National Advisory Council for Online Safety (Report of a National Survey of Children, their Parents and Adults regarding Online Safety). Retrieved from: **https://www.gov.ie/en/publication/ebe58-national-advisory-council-for-online-safety-nacos/**

Papatraianou, L. H., Levine, D., & West, D. (2014). Resilience in the face of cyberbullying: An ecological perspective on young people's experiences of online adversity. Pastoral Care in Education, 32(4), 264-283.

Pfattheicher, S., & Keller, J. (2015). The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European journal of social psychology*, *45*(5), 560-566.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. (2022). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*, 1-18.

Smith, P. K. (2016). Bullying: Definition, Types, Causes, Consequences, and Intervention. Social and Personality Psychology Compass, 10, 519-532.

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

Third, A., Collin, P., Fleming, C., Hanckel, B., Moody, L., Swist, T., & Theakstone, G. (2021). Governance, children's rights and digital health. Retrieved from: **https://www.governinghealthfutures2030.org/wp-content/uploads/2021/10/Governance-childrens-rights-and-digital-health.pdf**

Van Bommel, M., van Prooijen, J. W., Elffers, H., & Van Lange, P. A. (2012). Be aware to care: Public self-awareness leads to a reversal of the bystander effect. *Journal of Experimental Social Psychology*, *48*(4), 926-930.

Van Royen, K. V., Poels, K., Vandebosch, H., & Zaman, B. (2021). Think Twice to be Nice? A User Experience Study on a Reflective Interface to Reduce Cyber Harassment on Social Networking Sites. *International Journal of Bullying Prevention*, 1-12.

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.

Van Royen, K., Poels, K., & Vandebosch, H. (2016). Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. *Children and Youth Services Review*, *64*, 35-41

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a

systematic review: Garbage in, garbage out. *PloS one*, *15*(12), e0243300.

Williford, A., Elledge, L. C., Boulton, A. J., DePaolis, K. J., Little, T. D., & Salmivalli, C. (2013). Effects of the KiVa antibullying program on cyberbullying and cybervictimization frequency among Finnish youth. *Journal of Clinical Child & Adolescent Psychology*, *42*(6), 820-833.

Wu, J., Luan, S., & Raihani, N. (2022). Reward, punishment, and prosocial behavior: Recent developments and implications. *Current opinion in psychology*, *44*, 117-123.

Appendix

Table 1: Focus groups (FGs), sample structure

Focus groups	Number of participants	Sex	Age
FG1	9	Female	13-14
FG2	6	Female	16-17
FG3	8	Female	15-16
FG4	9	Female	15-16
FG5	6	Male	13-14
FG6	6	Male	15-16

Table 2: Interviews, sample structure

Sex and age	Number of interviews
Males, age 12	2 interviews
Males, age 13	1 interview
Males, age 14	1 interview
Males, age 15	1 interview
Males age 16	2 interviews
Females, age 12	1 interview
Females, age 13	1 interview
Females, age 14	1 interview
Females, age 15	3 interviews
Females, age 16	2 interviews

Acknowledgements

As an Elite-S fellow, Tijana Milosevic has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology grant number 13/RC/2106_P2.

Mr Derek Laffan also received funds from the Department of Education, Ireland, while working on the project.

We thank Prof. Sameer Hinduja and Ms. Anne Collier for kindly providing feedback on this report.

We would also like to thank Foróige for their help with participant recruitment. Special thanks to Ms. Angela Kinahan and ABC for all the administrative support which enabled the execution of this study; to Mr. Darran Heaney and Ms. Katrina Harrison for help with recruitment.

We thank Dushandthini Saravanan and Vivek Raj for their help with transcriptions.

We would like to thank Trill for allowing us to use their platform in our study and for their support.

Last but not least, our enormous gratitude goes to all the young people who took the time and energy to participate in this study, and we hope that this work will translate into meaningful policies made not just for but by them.

All photos of users used in demos are taken from community mockup templates available in Figma: https://www.figma.com/file/q3YK35FXUSrlZzjgd3NFqC/TikTok-UI-Screens-Community

The photos were published open access for TikTok by: https://www.figma.com/@pixsellz

and for Instagram by: https://www.figma.com/@arthurhazan

All the other photos are purchased and owned by the Anti-bullying Centre and do not require crediting.

Ionad Frithbhulaíochta
Anti-Bullying
Centre

ISBN Number: 978-1-911669-45-6

Contact details

tijana.milosevic@dcu.ie
@TiMilosevic