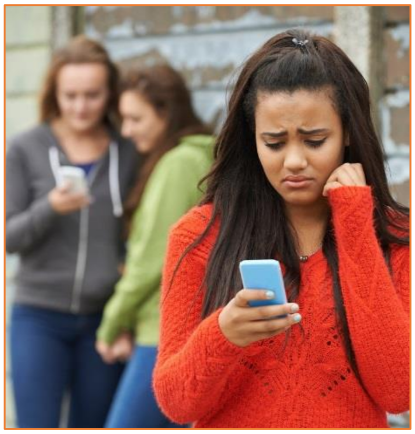
Media > Ethics > Initiative >

The University of Texas at Austin
Center for Media Engagement
Moody College of Communication

Filtering Out Cyberbullying:
The Ethics of Instagram's Anti-Bullying Filters



Highwaystarz / Onepixel.com / Modified

Of the many new ethical challenges raised by social media, one of the more worrisome is cyberbullying. A recent Pew survey found that 59% of American teens say they have experienced some type of cyberbullying online or on their cell phone. Bullying has been especially prominent on one of the most popular social media platforms, the image-intensive site Instagram. On October 9, 2018, Instagram announced that they would introduce a feature that utilizes machine learning techniques to detect bullying in images and captions posted on the platform. When detected, the images are reviewed by an Instagram employee to determine if they should be deleted. This new feature builds off a tool implemented last year that detects hurtful or offensive comments.

Instagram's new tools are intended to prevent bullying among users of its platform, a response to criticisms that it has failed to prevent bullying—in a recent study, for instance, "Instagram was highlighted as having become the vehicle most used for mean comments" (The Annual Bullying Survey 2017). Taylor Lorenz claims that "Teenagers have always been cruel to one another. But Instagram provides a uniquely powerful set of tools to do so" (2018). The head of Instagram, Adam Mosseri, hopes to limit the ability for users to utilize this "powerful set of tools" in a negative manner. In a blog post announcing the release of the new anti-bullying measures, he commented that they will "help us protect our youngest community members since teens experience higher rates of bullying online than others" (Mosseri, 2018). Many observers are hopeful that features like these will reduce the amount of negative posts and comments and help the many young Instagram users that struggle with bullying on the Internet.

As laudable as these goals are, the introduction of Instagram's anti-bullying measures has been meet with concern from many of its users. Some critics believe that these features are another instance of social networks limiting free speech on their sites for the purpose of making their platforms more desirable, thereby increasing the number of users and the site's profit (Blakely & Balaish, 2017). Many believe that companies like Facebook, who owns Instagram, have too much power when it comes to deciding who gets to say what on social media. In discussing the new anti-bullying filters, Kalev Leetaru argues that Silicon Valley is transitioning from the "early days of embracing freedom of speech at all costs" to becoming "the party of moderation and mindful censorship" (2018).

Critics of Instagram's latest moves are also concerned about the way these changes were planned and implemented. Instagram has not been fully transparent in revealing the techniques used to filter out negative images and comments: "no detail is given beyond that 'machine learning' is being used, no reassuring statistics on how much training data was used or the algorithm's accuracy rate when it entered service. Subsequent public statements are typically vague, claiming 'successes' or misleadingly worded statements that are not corrected when the media reports them wrong and rarely include any kind of accuracy statistics" (Leetaru, 2018). The lack of transparency, explicit standards, and accuracy data are a cause for concern. Due to this lack of transparency, many worry about implicit bias in the anti-bullying filters—or human moderators—that could eliminate posts which are not truly bullying in nature. However, Instagram argues that allowing external oversight into their filters and providing information on how they exactly work would allow for malicious users to "game" the system and avoid the anti-bullying controls.

Another worry centers on the issue of whether image content can be reliably identified as "bullying" in nature. For instance, some "split-screen" images compare a picture of a user to another photo in a negative fashion, but this isn't always the function of comparative collages. What makes an image—whether individually or as part of a collage—an essential part of an act of cyberbullying? Would posting or tagging an unflattering picture or undesirable image constitute bullying behavior? What about a user who posts an image mocking a public figure, politician, or famous celebrity? In more general terms, what differentiates a course of bullying from harsh criticism?

Instagram's new bullying filters could help to protect younger users from harmful attacks and hateful comments. However, these preventative measures come at a cost to the freedom of speech on the Internet and give more power and control of online content to large social media companies like Facebook. How much filtering do we need to purify our online ecosystems of cyberbullying?

Discussion Questions:

1. Should social media websites be held responsible for the content that is posted on their platforms by individual users?
2. How would you define cyberbullying? What features must a post have to count as cyberbullying? Can you think of another sort of non-bullying post that might have some or all of these features?
3. What ethical values are in conflict in Instagram's attempt to fight cyberbullying? How might you balance these values in dealing with cyberbullying?
4. What are the drawbacks to having human moderators determine which comments and posts are harmful and which should be allowed to remain? What are the drawbacks to machines or programs doing most of this work?

Further Information:

"The Annual Bullying Survey 2017: What 10,000 People Told Us About Bullying." *Ditch the Label*, July 2017. Available at: http://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2017/

Blakely, Jonathan, and Timor Balaish. "Is Instagram going too far to protect our feelings?" *CBS News*, August 14, 2017. Available at: http://www.cbsnews.com/news/inside-instagram/

Leetaru, Kalev. "Why Instagram's New Anti-Bullying Filter Is So Dangerous" *CNN Business*, October 11, 2018. Available at: http://www.forbes.com/sites/kalevleetaru/2018/10/11/why-instagrams-new-anti-bullying-filtering-is-so-dangerous/

Lorenz, Taylor. "Teens Are Being Bullied 'Constantly' on Instagram" *The Atlantic*, October 10, 2018. Available at: http://www.theatlantic.com/technology/archive/2018/10/teens-face-relentless-bullying-instagram/572164/

"A Majority of Teens Have Experienced Some Form of Cyberbullying." *Pew Research Center*, Washington, D.C. (September 17, 2018). Available at: http://www.pewinternet.org/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/

Mosseri, Adam. "New Tools to Limit Bullying and Spread Kindness on Instagram." October 9, 2018. Available at: http://instagram-press.com/blog/2018/10/09/new-tools-to-limit-bullying-and-spread-kindness-on-instagram/

Wakefield, Jane. "Instagram tops cyber-bullying study" *BBC News*, July 19, 2017. Available at: http://www.bbc.com/news/technology-40643904

Yurieff, Kaya. "Instagram says it will now detect bullying in photos." *Forbes*, October 9, 2018. Available at: https://www.cnn.com/2018/10/09/tech/instagram-anti-bullying-tools/

Authors:

Colin Frick & Scott R. Stroud, Ph.D.
Media Ethics Initiative
Center for Media Engagement
University of Texas at Austin
November 20, 2018