CENTER FOR DEMOCRACY & TECHNOLOGY

BRENNAN CENTER FOR JUSTICE

Social Media Monitoring in K-12 Schools:
Civil and Human Rights Concerns

Systems for monitoring students' social media posts are gaining popularity in K-12 education settings.[1] Some schools and school districts are turning to social media monitoring as a response to the threat of mass shootings.[2] Companies are marketing their social media monitoring services with claims that they can identify sexual content and drug and alcohol use; prevent mass violence, self-harm, and bullying; and/or flag students who may be struggling with academic or mental health issues and need help.[3] There is limited comprehensive data, but available figures, as well as statements from the companies themselves, suggest that spending by U.S. school districts on social media monitoring tools has risen substantially in recent years.[4] However, the claims of effectiveness by companies selling these products are largely unproven, and these tools can endanger the very students they are supposed to protect.[5] Surveilling students' social media activities raises the following serious privacy, free expression, and other civil and human rights concerns that schools, districts, and legislatures should safeguard against.[6]

1. Social media monitoring software for schools is experimental and has limited efficacy.

Social media monitoring techniques cannot live up to their claims of preventing violence.[7] Most social media monitoring software relies on algorithms to analyze posts. These tools are largely experimental and have significant technical limitations and accuracy problems.[8] Some of these tools rely on a predetermined "library" of words or phrases that could indicate potential risks of harm.[9] However, many words associated with harm (such as "bomb" or "shoot") are extremely common and have meanings that are entirely context-dependent. These types of tools will produce many false positives,[10] overwhelm schools with information, and subject far too many students to unnecessary surveillance given their limited efficacy.[11]

Machine learning, a type of artificial intelligence in which computers learn to identify patterns from examples, may capture slightly more context, but machine learning algorithms are still limited to recognizing patterns in the text that they have been trained on. They do not possess human-like abilities to interpret the meaning or the intent of the speaker.[12] Language use evolves quickly on social media—especially among young people—and novel uses of language are particularly likely to trip up an algorithm.[13] Studies show that social media monitoring algorithms tend to err in recognizing humor, sarcasm, and slang, all of which are common on social media.[14]

While human review is essential for interpreting human speech, even humans often err in understanding the meaning of social media posts. Age, gender, racial, and cultural differences can inhibit reviewers' ability to make sense of social media posts and magnify the risk of mistaking a joke for a serious threat.[15] These shortcomings suggest that the people best situated to accurately interpret others' posts, and identify possible risks, are peers and other community members, who are more likely to have a nuanced interpretation of their peers' posts.[16] However, school districts should tread carefully when encouraging students to report their peers' social media posts. Students' ability to form trusting and supportive relationships with their peers is critical to a healthy school environment.[17] Aggressive peer reporting systems risk undermining that trust.

2. Resources dedicated to social media monitoring may be best focused elsewhere.

Reliance on social media monitoring as a safety measure can divert much-needed resources from other school functions that could improve outcomes for all students.[18] Because of the limited utility of social media monitoring and the risks it poses to students' rights and health, diverting limited education funds to these systems could be detrimental to students and their families. Social media monitoring programs also put schools in the difficult position of having to monitor and respond to what students are saying off campus. This is a significant burden on schools that can detract from their educational mission.

3. Social media monitoring is not a reliable method to predict mass violence.

While tragic, mass shootings are statistically rare.[19] Because of the small sample size and the complexity of factors surrounding them, they cannot be reliably predicted, especially through automated tools. Any efforts to do so would flag a significant number of youth who pose little to no risk, causing lasting harm by subjecting innocent students to scrutiny from school authorities and potentially law enforcement.

4. Social media monitoring invades students' privacy.

Monitoring students' personal social media accounts can intrude upon their privacy, even when social media posts are viewable by the public or by a subset of people. Some monitoring systems are designed to send any post containing certain words or phrases directly to a school administrator.[20] While students may expect a teacher, principal, or parent to see an occasional public social media post, systematic monitoring is much more invasive. Systematic monitoring can reveal sensitive information about a student's personal life; for example, systems that monitor for sexual content may inadvertently out LGBTQ+ students. Often, the objective of social media monitoring is not only to surface what students are saying directly but to also infer information that they have not directly disclosed, such as their mental health status.[21]

Students face additional privacy risks when schools and districts create and store records noting that a social media post was flagged. Many jurisdictions lack policies about how long this data can be kept and how it can be used. Even if a student is ultimately cleared of any wrongdoing or the post turns out to be a false positive, they could still face negative repercussions if school records indicate that they were flagged and assessed. For example, if the data is shared with teachers or college admissions officers, it could impact how students are treated in the classroom or whether they get into college.

5. Social media monitoring can chill expressive activities that are critical for young people's development.

Reports show that online surveillance stifles expressive activities.[22] Because social media monitoring is targeted at surveilling speech, it risks dissuading students from expressing themselves, particularly when it comes to minority views or unpopular opinions. Because students use social media to connect with others, surveillance can also chill the development of meaningful relationships or engagement in political organizing.[23] Social media monitoring programs often focus on detecting possible mental health issues, which could have the counterproductive effect of discouraging students from openly discussing mental health or reaching out for support on social media. Young people are particularly vulnerable to chilling effects, given the unequal balance of power between themselves and authority figures at school and elsewhere, and they need breathing space to explore new ideas, develop their sense of selves, and learn to participate in a civic society.

6. Social media monitoring can disproportionately burden minority, underserved, or vulnerable students.

Overbroad surveillance and harmful chilling effects from social media monitoring are likely to have a disproportionate impact on students from minority or vulnerable communities, including students of color, immigrant students, students with disabilities, and Muslim students or students affiliated with another religious minority.

In the U.S., for instance, students of color are punished at higher rates than white students and are punished more severely for less serious infractions; girls of color and students with disabilities—particularly those who are also of color—are especially likely to experience disproportionate punishment.[24] These dynamics are likely to play out in social media monitoring as well. Even if software manages to flag white students and students of color at equal rates, students of color and Muslim students are at a higher risk of being punished or subjected to law enforcement contact based on a flagged post.[25]

In addition, social media monitoring tools tend to have lower accuracy rates for minority speakers, including those who post in languages other than English and those who use vernacular associated with a subgroup.[26] Even when posts are reviewed by humans, majority reviewers are more likely to misunderstand the meaning of posts by minority speakers.[27]

Finally, the chilling effects from social media monitoring may also be amplified for marginalized communities. Students who are already at a higher risk for surveillance, punishment, and law enforcement contact may be particularly wary of monitoring and may censor themselves more.[28]

Endnotes

1. Faiza Patel, Rachel Levinson-Waldman, Jun Lei Lee, and Sophia DenUyl, "School Surveillance Zone," *The Brennan Center for Justice*, April 20, 2019, https://www.brennancenter.org/analysis/school-surveillance-zone; Tom Simonite, "Schools are Mining Students' Social Media Posts for Signs of Trouble," *Wired*, August 20, 2018, https://www.wired.com/story/algorithms-monitor-student-social-media-posts/.
2. *Ibid*
3. "Bark for Schools: Student and School Safety for G Suite and Office 365," last modified 2019, https://www.bark.us/schools; Social Sentinel, last modified 2019, https://www.socialsentinel.com/.
4. Patel *et al.*, "School Surveillance Zone"; "Social Sentinel's Year End Look at the Impact of Technology and School Safety," *Social Sentinel*, last modified January 16, 2019, https://www.socialsentinel.com/social-sentinel's-year-end-look-at-the-impact-of-technology-and-school-safety.
5. "Social Sentinel's Year End Look at the Impact of Technology and School Safety"; Faiza Patel and Rachel Levinson-Waldman, "Monitoring Kids' Social Media Accounts Won't Prevent the Next School Shooting," *Washington Post* (Washington, DC), Mar. 5, 2018, https://beta.washingtonpost.com/news/posteverything/wp/2018/03/05/monitoring-kids-social-media-accounts-wont-prevent-the-next-school-shooting/.
6. "Technological School Safety Initiatives: Considerations to Protect All Students," *Center for Democracy and Technology and Brennan Center for Justice,* June 6, 2019, https://cdt.org/insight/technological-school-safety-initiatives-considerations-to-protect-all-students/.
7. Aaron Leibowitz, "Could Monitoring Students on Social Media Stop the Next School Shooting?" *New York Times* (New York, NY), Sep. 6, 2018, https://www.nytimes.com/2018/09/06/us/social-media-monitoring-school-shootings.html; Brian Resnick and Javier Zarracina, "This Cartoon Explains Why Predicting a Mass Shooting is Impossible," *Vox*, August 5, 2019, https://www.vox.com/science-and-health/2018/2/22/17041080/predict-mass-shooting-warning-sign-research; Bruce Schneider, "Why Mass Surveillance Can't, Won't, and Never Has Stopped A Terrorist," *Digg*, March 11, 2015, "https://digg.com/2015/why-mass-surveillance-cant-wont-and-never-has-stopped-a-terrorist.
8. Natasha Duarte, Emma Llanso, and Anna Loup, "Mixed Messages? The Limits of Automated Social Media Content Analysis," *Center for Democracy and Technology*, November 28, 2017, https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/.
9. "Supporting Effective Safeguarding and Child Protection Practices," *Impero EdAware*, 2018, https://kc0eiuhlnmqwdxy1ylzte9ii-wpengine.netdna-ssl.com/us/wp-content/uploads/sites/16/2018/12/Impero-EdAware-US-CLOUD-2019.pdf; Social Sentinel; Security Sales & Integration Staff, "Rekor Releases Solution for Early Warning Threat Detection, Identification in Schools," *Security Sales & Integration*, August 6, 2019, https://www.securitysales.com/fire-intrusion/rekor-threat-detection-schools/.
10. Leibowitz, "Could Monitoring Students on Social Media Stop the Next School Shooting?"
11. "Algorithmic Systems in Education: Incorporating Equity and Fairness When Using Student Data," *Center for Democracy and Technology*, August 12, 2019, https://cdt.org/insight/algorithmic-systems-in-education-incorporating-equity-and-fairness-when-using-student-data/.
12. Duarte *et al.*, "Mixed Messages? The Limits of Automated Social Media Content Analysis."
13. *Ibid.*
14. *Ibid.*
15. danah boyd, "For Privacy, Teens Use Encoded Messages Online," *Science Friday*, February 27, 2014, https://www.sciencefriday.com/articles/for-privacy-teens-use-encoded-messages-online/; Alex Hern, "Facebook Translates 'Good Morning' into 'Attack Them', Leading to Arrest," *Guardian* (London, UK), Oct. 24, 2017, https://www.theguardian.com/technology/2017/oct/24/facebook- palestine-israel-translates-good-morning-attack-them-arrest.
16. Desmond Upton Patton, Philipp Blandfort, William R. Frey, Michael B. Gaskell, and Svebor Karaman, "Annotating Twitter Data from Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators," *ResearchGate*, January 2019,

https://www.researchgate.net/publication/330261697_Annotating_Twitter_Data_from_Vulnerable_Populations_Evaluating_Disagreement_Between_Domain_Experts_and_Graduate_Student_Annotators.
17. Natasha Duarte, "Six Considerations Missing from the School Safety and Data Conversation," *Center for Democracy and Technology*, March 13, 2019, https://cdt.org/blog/six-considerations-missing-from-the-school-safety-and-data-conversation/.
18. Patel *et al*, "School Surveillance Zone."
19. Maggie Koerth-Baker, "Mass Shootings Are a Bad Way to Understand Gun Violence," *FiveThirtyEight*, October 3, 2017, https://fivethirtyeight.com/features/mass-shootings-are-a-bad-way-to-understand-gun-violence/; "School Violence Myths," *University of Virginia Curry School of Education and Human Development*, July 26, 2015, https://curry.virginia.edu/faculty-research/centers-labs-projects/research-labs/youth-violence-project/violence-schools-and-5 (see Myth 5, noting that while "[g]un violence is a serious problem in the United States … this problem is less likely to occur in schools than almost any other location").
20. "Bark for Schools: Student and School Safety for G Suite and Office 365"; Social Sentinel.
21. Natasha Singer, "In Screening for Suicide Risk, Facebook Takes on a Tricky Public Health Role," *New York Times* (New York, NY), Dec. 31, 2018, https://www.nytimes.com/2018/12/31/technology/facebook-suicide-screening-algorithm.html.
22. Kaveh Waddell, "How Surveillance Stifles Dissent on the Internet," *Atlantic*, April 5, 2019, https://www.theatlantic.com/technology/archive/2016/04/how-surveillance-mutes-dissent-on-the-internet/476955/.
23. Emily Witt, "From Parkland to Sunrise: A Year of Extraordinary Youth Activism," *New Yorker*, February 13, 2019, https://www.newyorker.com/news/news-desk/from-parkland-to-sunrise-a-year-of-extraordinary-youth-activism.
24. "We Came to Learn: A Call to Action for Police-Free Schools," *Advancement Project*, last modified 2019, https://advancementproject.org/wecametolearn/; "2015-16 Civil Rights Data Collection: School Climate and Safety," *U.S. Department of Education and Office of Civil Rights*, April 2018, last revised May 2019, https://www2.ed.gov/about/offices/list/ocr/docs/school-climate-and-safety.pdf; "Black Girls Matter: Pushed Out, Overpoliced, and Underprotected," *African American Policy Forum*, February 4, 2015, http://aapf.org/recent/2014/12/coming-soon-blackgirlsmatter-pushed-out-overpoliced-and-underprotected; Monica Rhor, "Pushed Out and Punished: One Woman's Story Shows How Systems are Failing Black Girls," *USA Today*, May 15, 2019, https://www.usatoday.com/in-depth/news/nation/2019/05/13/racism-black-girls-school-discipline-juvenile-court-system-child-abuse-incarceration/3434742002/ ; Libby Nelson, "The Hidden Racism of School Discipline, in 7 Charts," *Vox*, October 31, 2015, https://www.vox.com/2015/10/31/9646504/discipline-race-charts; Darby Derrick and John L. Rury, "When Black Children Are Targeted for Punishment," *New York Times* (New York, NY), Sep. 25, 2017, https://www.nytimes.com/2017/09/25/opinion/black-students-little-rock-punishment.html; Evie Blad and Alex Harwin, "Analysis Reveals Racial Disparities in School Arrests," *PBS NewsHour*, February 27, 2017, https://www.pbs.org/newshour/education/analysis-reveals-racial-disparities-school-arrests.
25. DallasNews Administrator, "Ahmed Mohamed Swept Up, 'Hoax Bomb' Charges Swept Away as Irving Teen's Story Floods Social Media," *Dallas Morning News* (Dallas, TX), Sep. 15, 2015, https://www.dallasnews.com/news/2015/09/16/ahmed-mohamed-swept-up-hoax-bomb-charges-swept-away-as-irving-teen-s-story-floods-social-media/.
26. Duarte *et al.*, "Mixed Messages? The Limits of Automated Social Media Content Analysis"; Shirin Ghaffary, "The Algorithms that Detect Hate Speech Online are Biased Against Black People," *Vox*, August 15, 2019, https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter?stream=future.
27. Leila Ettachfini, "Court Reporters May Be Writing Down Black People's Testimonies Wrong," *Vice News*, May 23, 2019, https://www.vice.com/en_us/article/ywynzj/court-reporters-write-down-black-testimonies-wrong-study; John Eligon, "Speaking Black Dialect in Courtrooms Can Have Striking Consequences," *New York Times* (New York, NY), Jan. 25, 2019, https://www.nytimes.com/2019/01/25/us/black-dialect-courtrooms.html.
28. Sarah Brayne, "Surveillance and System Avoidance: Criminal Justice Contact and Institutional Attachment," *American Sociological Review* 79, no. 3 (2014): 367-391, accessed September 24, 2019, https://journals.sagepub.com/doi/10.1177/0003122414530398.