Social Media Suicide Watch: Privacy vs. Beneficence

DOI: 10.52214/vib.v11i.14035

Katherine Prothro*

Abstract

Two tragedies illustrate the need to both monitor suicidal ideation more effectively and protect individuals from harmful influences. In the UK, a 14-year-old girl named Molly Russel died by suicide after consuming self-harming content online, while her own distressing signals on social media went unnoticed. In contrast, Conrad Roy III died by suicide after encouragement from his girlfriend via texts and calls, demonstrating the devastating consequences of alerting the wrong person to a mental health crisis. These cases highlight the moral ambiguity of tracking public expressions of suicidal ideation – when is it appropriate to monitor this data, and with whom should it be shared? The public was compelled to weigh in on this issue in 2014 when Samaritans implemented an app that scanned Twitter users' posts for signs of suicidal ideation and alerted fellow users. There was an immediate backlash due to concerns over privacy and the potential for stalkers and bullies to misuse this data and encourage suicide or self-harm, like Roy's girlfriend did. However, for some individuals, like Molly Russel, social media posts may be a cry for help, which could be met with thoughtfully crafted interventions. This paper proposes a new model for monitoring suicidal ideation online – one that promotes autonomy, respects confidentiality, and provides evidence-based support. This model includes direct user alerts, safety planning, and opt-in data-sharing with mental health professionals.

Keywords: Suicide, Social Media, Suicidal Ideation, Privacy, Autonomy, Opt-in Data Sharing, Mental Health

Introduction

Molly Russel is a household name in the UK. After dying by suicide at age 14, an inquiry into her death revealed that Molly was suffering from depression and the negative effects of harmful online content. She was repeatedly exposed to posts that displayed hopelessness and misery. Her own Twitter account revealed signals of her worsening mental health crisis.[1] While better content censoring could have helped, an algorithm capable of detecting her worsening

[1] Crawford, A. (2023, November 29). Molly Russell: Tech firms still failing after teenager's death, says father. *BBC News*. https://www.bbc.com/news/uk-67556756

* Katherine Prothro, medical student at Stanford University, editor for *Voices in Bioethics*.

mental state and alerting trusted individuals could have also served as a crucial safety net. Notifying mental health professionals and law enforcement of potential suicide risk is considered a protective factor.

Yet, this same technology could have unintended and devastating consequences by alerting a wrongdoer. In 2014, Conrad Roy III died by suicide after his girlfriend encouraged him to end his life via text messages. While this was not a social media case, after he filled his truck with toxic fumes, she urged him to get back inside and listened as he died, making no attempt to intervene.[22] This case highlights the danger of alerting the wrong person to suicidal ideation. When such sensitive information falls into the wrong hands, it can be weaponized.

These two opposing stories emphasize the double-edged nature of using technology to detect and respond to mental health crises. So, who do we save—Molly or Conrad? Which scenario is more likely? The public weighed in on this dilemma in 2014, when the charity Samaritans launched a Twitter app that allowed users to receive alerts if someone they followed posted messages suggesting suicidal thoughts.[3] There was almost immediate backlash. Critics flagged concerns about privacy, surveillance, and the potential for harm. Yet, the question remains: While some posts are publicly visible, others are deliberately restricted to followers only. Flagging the former may raise ethical questions, but alerting others to the second more clearly violates the app's own terms of use. Could an app truly prevent suicide and promote well-being? And if so, does that outcome justify infringing on individuals' expectations about privacy or the nonconsensual analysis of their posts?

Ethical Analysis

The strongest argument in favor of apps like Samaritan's Radar is their potential to predict individuals at risk for suicide and self-harm accurately. Psychiatrists and most primary care physicians are trained to assess suicide risk based on factors such as previous suicide attempts, chronic pain, legal problems, and financial difficulties.[4] However, these traditional methods are often too broad to be consistently effective in clinical settings.[5] Apps like Samaritan's Radar could refine risk prediction by providing real-time data, potentially improving intervention efforts. AI can be used to retrospectively analyze posts from individuals who died by suicide to identify patterns imperceptible to the human eye. These algorithms could then be applied to current users to more accurately identify at-risks populations.

However, even if an app offers the possibility of more accurately identifying suicide risk, it may be that therapists and clinicians are better trained to assess the sincerity of flagged posts than an app. For example, a post stating, *"I have*

[2] Seelye, K. Q. (2017, June 16). Michelle Carter is found guilty in texting suicide case. *The New York Times*. https://www.nytimes.com/2017/06/16/us/suicide-texting-trial-michelle-carter-conrad-roy.html

[3] Orme, J. (2014, November 7). Samaritans pulls 'suicide watch' Radar app over privacy concerns. *The Guardian*. https://www.theguardian.com/society/2014/nov/07/samaritans-radar-app-suicide-watch-privacy-twitter-users

[4] Brown, G. K., & Green, K. L. (2014). A review of evidence-based follow-up care for suicide prevention: Where do we go from here? *American Journal of Preventive Medicine*, 47(3 Suppl 2), S209–S215. https://doi.org/10.1016/j.amepre.2014.06.006; Centers for Disease Control and Prevention. (2024, April 25). *Risk and protective factors for suicide*. U.S. Department of Health and Human Services. https://www.cdc.gov/suicide/risk-factors/index.html

[5] Hawton, K., Lascelles, K., Pitman, A., Gilbert, S., & Silverman, M. (2022). Assessment of suicide risk in mental health practice: Shifting from prediction to therapeutic assessment, formulation, and risk management. *The Lancet Psychiatry*, 9(6), 518–528. https://doi.org/10.1016/S2215-0366(22)00232-2; Lucey, J. V., & Matti, B. (2021). Suicide risk assessment: Time to think again? *Irish Journal of Psychological Medicine*, 40(3), 323–325. https://doi.org/10.1017/ipm.2021.76

so much homework, I'm gonna kill myself," may warrant attention but more likely reflects a state of frustration rather than real suicidal intent.

Furthermore, apps may struggle to recognize the nuanced stages of suicide risk, such as contemplation, suicidal intent, the development of a plan, intent to act, and suicidal behaviors. These distinctions are critical in determining the appropriate level of intervention, something that human judgment may handle more effectively than algorithmic detection.

Additionally, existing systems designed to prevent suicide already infringe on individual rights. For example, under HIPAA, therapists and psychiatrists are permitted to contact emergency services, for example by calling 911, when there is a threat to safety.[6] Psychiatric hospitalization can help prevent suicidal behavior by providing immediate safety and stabilization. It also offers an opportunity to initiate treatment, including medication and counseling. Suicide safety planning and follow up contact initiated during these hospitalizations has also been shown to improve suicidal ideation and psychological stress in the months following discharge.[7]Alerting trusted individuals or emergency services is one of the most effective ways to intervene and save lives. For many families who have lost a relative to suicide, the opportunity to protect a life far outweighs the concerns about the temporary intrusion of privacy.

Regarding the issue of consent, patients who disclose suicidal ideation to a healthcare professional typically do so with the understanding that providers have a mandatory reporting policy. They are required to share this stipulation with patients at the outset. In the context of a psychiatric or mental healthcare appointment, patients share their inner thoughts knowing their exact audience. However, the Samaritan's Radar app did not notify users that an algorithm was analyzing their content for signs of distress. Even more concerning, email alerts were sent to anyone who signed up to follow a given user, without that user's consent. Some may argue that the posts on Twitter are public and can be viewed by anyone. However, this does not equate to consent for an app to actively warn or alert others about a user's mental health status. Doing so risks breaching privacy and may misrepresent the user's state.

Furthermore, alerting users whom the individual did not designate as trusted without consent risks further harm. Their relationship to the at-risk user would be unclear, and their response could inadvertently worsen the situation. They may lack the appropriate words to offer support. In worse cases, they could unintentionally shame, stigmatize, or even bully the individual. When Conrad Roy shared his suicidal ideation with his girlfriend, she weaponized that vulnerability. Similarly, individuals on Samaritan's Radar may deliberately follow trends of emotional distress to hurt, rather than help individuals.

Additionally, such alerts might discourage people from posting altogether, robbing friends and relatives of the opportunity to naturally reach out in a way that feels organic rather than intrusive. In the best-case scenario, those who see a concerning post could offer resources or, if necessary, contact emergency services on their own initiative.

Rethinking Samaritan's Radar: A Path Forward

To balance user privacy with the need to protect individuals struggling with mental health issues, apps should be able to analyze publicly posted messages for signs of suicidal ideation. However, rather than notifying others without the user's consent, the app should first alert the individual directly, presenting the detected concerns. It should then offer

[6] ((45 CFR §164.512(j)).

[7] Stanley, B., Brown, G. K., Brenner, L. A., Galfalvy, H. C., Currier, G. W., Knox, K. L., Chaudhury, S. R., Bush, A. L., & Green, K. L. (2018). Comparison of the Safety Planning Intervention with follow-up vs usual care of suicidal patients treated in the emergency department. *JAMA Psychiatry*, 75(9), 894–900. https://doi.org/10.1001/jamapsychiatry.2018.1776

options such as calling a suicide helpline, contacting emergency services, or reaching out to a trusted person of their choice (e.g., a parent, friend, or designated support). This approach provides support to users while reinforcing their agency by giving them choice in how to proceed.

Even more effective would be the inclusion of a suicide safety plan, a tool that has been shown to reduce suicidal behavior significantly.[8] If users have previously and voluntarily identified themselves as high risk, the app could provide this plan proactively. Otherwise, it could be offered in moments of acute crisis, delivering evidence-based support at the time it is most needed.

It may also be beneficial to have an opt-in feature that allows apps to share information with trusted health professionals or guardians. With user permission, the app could notify psychiatrists, not just of suicidal content, but of posts suggesting worsening mental stability, enabling earlier intervention before it's too late. If the content raises serious concern, and the user has given prior consent, the psychiatrist may contact a parent or an appropriate authority to ensure the individual's immediate safety.

Furthermore, app developers should collaborate with trained psychiatrists when designing these technologies, given their expertise in identifying and treating suicidal behavior. They may provide insight into crisis management strategies that may not be familiar to those outside of the mental health field. Before implementation, such apps should be examined through formal studies to assess their effectiveness and psychological impact. To justify their use, it is essential to prove that they reduce suicide attempts and cases.

While the Samaritan app was created with good intentions, it overlooked important ethical principles. It alerted the wrong people and bypassed users entirely, failing to respect their autonomy or offer them resources. However, one app should not ruin all future progress. Countless people express despair, hopelessness, and signs of self-harm and suicidal ideation online. With careful design, we don't have to choose between saving Molly or protecting Conrad. We can help them both. It is time we revisit the issue of using technology as a tool for suicide prevention.

[8] Orme, J. (2014, November 7). Samaritans pulls 'suicide watch' Radar app over privacy concerns. *The Guardian*. https://www.theguardian.com/society/2014/nov/07/samaritans-radar-app-suicide-watch-privacy-twitter-users