Philosophy & Technology (2018) 31:669–684
https://doi.org/10.1007/s13347-018-0336-0

COMMENTARY

Ethics and Artificial Intelligence: Suicide Prevention on Facebook

Norberto Nuno Gomes de Andrade[1] · Dave Pawson[1] · Dan Muriello[1] · Lizzy Donahue[1] · Jennifer Guadagno[1]

Received: 1 October 2018 / Accepted: 16 November 2018 / Published online: 26 December 2018

Abstract

There is a death by suicide in the world every 40 seconds, and suicide is the second leading cause of death for 15–29-year-olds. Experts say that one of the best ways to prevent suicide is for those in distress to hear from people who care about them. Facebook is in a unique position—through its support for networks and friendships on the site—to help connect a person in these difficult situations with people who can support them. Connecting people with the resources they need is part of Facebook's ongoing efforts to help build a safe community inside and outside of Facebook. This article provides a brief overview of how Facebook's work to develop suicide prevention tools started and evolved, and the ethical considerations which surfaced during the process in the form of concrete product decisions around the implementation of these tools. This article is structured into three sections. Section 1 reviews what has been done in this space and lists and briefly describes other suicide prevention apps and tools. Section 2 describes Facebook's overall approach to suicide prevention. Here, we'll delve first into how that approach originated and how it was influenced by the external community's proactive interactions with Facebook, highlighting our unique position to help address the problem. Afterwards, we'll explain how that approach evolved, describing its various stages and iterations: understanding, reactive reporting, queue prioritization, and proactive reporting. This section describes the tools and resources Facebook has developed for people who may be at risk. Particular attention is devoted to the use of ArtificiaI Intelligence (AI) and Machine Learning (ML) to detect posts or live videos where someone might be expressing thoughts of suicide. Section 3 will elaborate on the ethical questions addressed when developing our approach and when making concrete product decisions to implement our suicide prevention tools. In this last section, we'll expound the competing values and interests that were at stake during the product development process, and how we reached ethical balances between them.

Keywords Facebook · Suicide · Ethics · ArtificiaI Intelligence AI · Machine Learning

✉ Norberto Nuno Gomes de Andrade
nandrade@fb.com

[1] Facebook, Menlo Park, California, USA

1 Introduction

As estimated by the World Health Organization (WHO), every year close to 800,000 people die from suicide, which is one person every 40 seconds.[1] Beyond actual deaths by suicide, there are indications that for each adult who died of suicide there may have been more than 20 others attempting suicide.[2] In addition, it is estimated that at least six people are directly affected by each suicide death.[3]

Suicide is a global phenomenon. In fact, 78% of suicides occurred in low- and middle-income countries in 2015.[4] Suicide accounted for 1.4% of all deaths worldwide, making it the 17th leading cause of death in 2015.[5] The WHO further reports that in the last 45 years suicide rates have increased by 60% worldwide.[6] Suicide is now among the three leading causes of death among those aged 15–44 (male and female).[7] And youth suicide is increasing at the fastest rate.[8]

In the USA, nearly 45,000 suicides occurred in 2016—more than twice the number of homicides—making it the 10th leading cause of death.[9] More than half (54%) of people who died by suicide did not have a known mental health condition.[10] A nationwide survey of high school students in the USA found that 16% of students reported seriously considering suicide, 13% reported creating a plan, and 8% reported trying to take their own life in the 12 months preceding the survey.[11] Each year, approximately 157,000 youth between the ages of 10 and 24 are treated at emergency departments across the USA for self-inflicted injuries.[12]

Over the past 10 years, Facebook has worked with institutions and experts working in the field of suicide prevention, namely the National Suicide Prevention Lifeline, Crisis Text Line, and Forefront, to discuss ways to address this problem. Experts have repeatedly stated that one of the best ways to prevent suicide is for those in distress to hear from people who care about them, and that Facebook is in a unique position—through the friendships and connections that are created and fostered on our platform—to help connect a person in distress with people who can support them.[13] The investment and commitment to help support these connections is part of Facebook's ongoing effort to help build a safe community on and off Facebook.[14]

[1] http://www.who.int/news-room/fact-sheets/detail/suicide, http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/
[2] *ibid*
[3] Preventing suicide: a resource for media professionals, update 2017. Geneva: World Health Organization; 2017 (WHO/MSD/MER/17.5). Licence: CC BY-NC-SA 3.0 IGO. http://apps.who.int/iris/bitstream/handle/10665/258814/WHO-MSD-MER-17.5-eng.pdf;jsessionid=AC6E8947F094647D8B6B5986641C8653?sequence=1
[4] http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/
[5] *ibid*
[6] https://www.befrienders.org/suicide-statistics
[7] *ibid*
[8] *ibid*
[9] https://www.washingtonpost.com/news/to-your-health/wp/2018/06/07/u-s-suicide-rates-rise-sharply-across-the-country-new-report-shows/?noredirect=on&utm_term=.c4c2a0acf70a
[10] https://www.cdc.gov/vitalsigns/suicide/
[11] https://www.cdc.gov/healthcommunication/toolstemplates/entertainmented/tips/SuicideYouth.html
[12] *ibid*
[13] https://www.facebook.com/fbsafety/videos/1098099763561194/
[14] https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/

This article provides a brief overview of how Facebook's work to develop suicide prevention tools started and evolved, and the ethical considerations which surfaced during the process in the form of concrete product decisions around the implementation of these tools.

This article is structured into three sections. Section 1 reviews what has been done in this space and lists and briefly describes other suicide prevention apps and tools. Section 2 describes Facebook's overall approach to suicide prevention. Here, we'll delve first into how that approach originated and how it was influenced by the external community's proactive interactions with Facebook, highlighting our unique position to help address the problem. Afterwards, we'll explain how that approach evolved, describing its various stages and iterations: understanding, reactive reporting, queue prioritization, and proactive reporting. This section describes the tools and resources Facebook has developed for people who may be at risk. Particular attention is devoted to the use of ArtificiaI Intelligence (AI) and Machine Learning (ML) to detect posts or live videos where someone might be expressing thoughts of suicide. Section 3 will elaborate on the ethical questions addressed when developing our approach and when making concrete product decisions to implement our suicide prevention tools. In this last section, we'll expound the competing values and interests that were at stake during the product development process, and how we reached ethical balances between them.

2 Suicide Prevention: Brief Review of the Work Done in This Area

The prevalence of suicide worldwide in conjunction with the difficulty in receiving specialized support and care has led to an emergence of interest in using technology to facilitate care in this space. The recent movement of phone-based suicide support providers towards use of SMS text messaging and other internet chat services is one example (National Suicide Prevention Lifeline,[15] Crisis Text Line,[16] Talkspace[17]). These services seek to lower barriers that separate people in need from receiving care.

Digital assistant product makers such as Apple (Siri),[18] Google Assistant,[19] and Amazon (Echo/Alexa[20]) have also stepped into the field by augmenting their services to direct people towards suicide prevention[21] resources after receiving clear verbal indication of suicidal thoughts. The social networking space is not without its own prior work. While these services provide clear value for people suffering from suicidal ideation, they do not use novel detection techniques to go beyond simple keyword matching. The UK-based non-profit organization Samaritans briefly provided a product called Radar,[22] which also used keyword matching, described as "a free web application that monitors your friends' Tweets, alerting you if it spots anyone who may be struggling to cope".[23]

[15] https://suicidepreventionlifeline.org/
[16] https://www.crisistextline.org/
[17] https://www.talkspace.com/
[18] https://www.apple.com/ios/siri/
[19] https://assistant.google.com/
[20] https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/
[21] https://med.stanford.edu/news/all-news/2016/03/hello-siri-im-depressed.html
[22] https://www.samaritans.org/news/samaritans-launches-twitter-app-help-identify-vulnerable-people
[23] Radar parked widespread criticism based on the fact that it made concerning Tweets known to the author's followers. Radar was deactivated less than a month after its launch.

Recent academic studies have demonstrated promising attempts at utilizing Machine Learning to predict risk of suicide based on information in medical records,[24] though this sort of technique has not yet been applied at scale in the medical community.

The following list comprises some of the most relevant apps, services, and research initiatives developed for or containing features aimed at helping prevent suicide:

- Radar[25]

– Radar, as mentioned above, was an app developed by the UK non-profit Samaritans which allowed Twitter users to opt in to receiving notifications when a person they follow posts a tweet which the Radar system flags as possibly suicidal via an AI algorithm.

- Woebot[26]

– Woebot is an automated conversational agent (chat bot), built based on research in cognitive behavioral therapy. It's available either through Facebook's Messenger platform or via anonymous mobile apps for iOS or Android, and is driven by natural language processing. It provides a lightweight therapeutic service which has been shown to reduce anxiety and depression among test participants.

- Analysis of medical records[27]

– Researchers from Vanderbilt University recently published a study in which they trained a Machine Learning system to predict suicidal risk based on medical record data.

- Siri/Google Assistant/Alexa/Cortana

– These are voice assistant services which are built into popular computing platforms (iOS, Android, Amazon Echo, Windows). To varying degrees, they have features which direct people to suicide prevention resources based on a limited set of pre-programmed trigger words and phrases. Siri, for example, responds to the phrase "I want to kill myself" with: "It sounds like talking to someone might help. The National Suicide Prevention Lifeline offers confidential, one-on-one support, 24 hours a day" and offers to dial the number.

[24] http://journals.sagepub.com/doi/abs/10.1177/2167702617691560
[25] https://www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar
[26] https://www.woebot.io/; https://mental.jmir.org/2017/2/e19/
[27] http://journals.sagepub.com/doi/abs/10.1177/2167702617691560

- Suicide Prevention Lifeline[28]/Crisis Text Line[29]/Talkspace[30]

– These are services which connect people with either volunteer or professionally trained crisis support representatives via a chat interface.

- SAM[31]

– SAM is an anxiety management app developed in collaboration with researchers at UWE, Bristol. It provides a tracking interface and customized resources and exercises to reduce anxiety.

- Cogito Companion[32]

– This is a mood tracking app which is based on analysis of audio diary entries.

3 Facebook's Approach to Suicide Prevention: Team and Evolution

As far back as 10 years ago, Facebook felt strongly that our reach and resources put us in a unique position to help with certain social and personal issues that people in our platform may be going through. As an example, we started building thoughtful functionality to better manage the difficult situation that friends and family endure when their loved ones pass away. In those situations, and as a way to help handle the account safely and sensitively after the account owner's death, we offer designated friends the limited capability to manage the account of the deceased ones in their absence.[33] Another example of how we help people feel supported when specific challenging situations arise is our breakup management tool. This feature enables people to better manage a difficult romantic breakup experience by offering specific tools managing how they interact with former loves, helping people end relationships on Facebook with greater ease, comfort, and sense of control.[34] The suicide prevention tools we explain in the remainder of this paper are another example of the resources we are developing to help people who may be going through difficult moments in their lives.[35]

[28] https://suicidepreventionlifeline.org/
[29] https://www.crisistextline.org/
[30] https://www.talkspace.com/
[31] http://sam-app.org.uk/
[32] https://itunes.apple.com/us/app/cogito-companion/id1159040768?mt=8; https://play.google.com/store/apps/details?id=com.cogito.companion&hl=en
[33] https://newsroom.fb.com/news/2015/02/adding-a-legacy-contact/
[34] https://newsroom.fb.com/news/2015/11/improving-the-experience-when-relationships-end/
[35] These efforts are aligned with many other initiatives that we are currently pursuing, namely in the areas of wellbeing and accessibility. https://newsroom.fb.com/news/2017/12/hard-questions-is-spending-time-on-social-media-bad-for-us/; https://newsroom.fb.com/news/2018/08/manage-your-time/; https://www.wired.com/2015/02/meet-team-makes-possible-blind-use-facebook/

3.1 Understanding

Suicide prevention tools have been available on Facebook for more than 10 years. We developed these tools in collaboration with mental health organizations including save.org, Forefront Suicide Prevention, and National Suicide Prevention Lifeline, and with input from people who have personal experience thinking about or attempting suicide.[36] As such, first and foremost, we approached the problem by collaborating with experts in the field. Our expertise at building social networking and scalable software systems in no way qualified us to reinvent suicide prevention, so we relied heavily on academics, non-profit organizations and experts to share their domain knowledge with us, and to help us determine how to best use our technology to apply their learnings. Our first step in addressing the complex and widespread problem of suicide was to learn and understand from those that have thoroughly studied it, from those with expertise working on preventative measures, and from those who have experienced attempting suicide.

These resources were invaluable for us to understand the latest approaches in this area, along with its ethical considerations. Over time, and as we developed and rolled out specific tools in this area, we continued to collaborate with external experts, tapping into their knowledge and seeking their critique of the experiences we build.

3.2 Reactive Reporting

Based on what we learned from our research and collaboration, our early efforts in suicide prevention were focused on reactive response. A person posts something that could be seen as demonstrating suicidal ideation; and a friend who sees the post is offered a number of reactive options. In much the same way that any post can be reported for violating our community standards, a post can be reported for seemingly demonstrating a risk of suicide or self-harm. If someone goes down the path of making such a report, they are presented with options in the reporting flow. They may reach out to the person in distress directly. Through the option of connecting with mutual friends, they may speak with their own friends to make sure they are properly supported. And ultimately, they can choose to submit a report and ask for help. When such a report is submitted, it undergoes human review at Facebook, and one of three outcomes takes place:

1. The post is deemed not to be about suicide. In this case, the report is closed, and no action is taken.[37]
2. The post is deemed to demonstrate potential suicidal intent. The next time the person in distress returns to Facebook, they are presented with options to help them

[36] In particular, we worked extensively with Dr. Dan Reidenberg (SAVE.org non-profit), acknowledged expert in the field. We also worked closely with members of Forefront Suicide Prevention, a Center for Excellence at University of Washington, and with leaders at National Suicide Prevention Lifeline. We also met with and learned from experts in suicide prevention field from around the world, including leaders in India, Japan, Vietnam, Thailand, and the UK. Furthermore, we conducted multiple rounds of interviews with people who have had past experiences with suicide attempts or serious thoughts of suicide, in addition to talking with friends and family members of those who died by suicide or attempted suicide.

[37] If the report is not about suicide but violates our community standards, the report will be reviewed for that particular violation.

manage their situation. Use of all such resources is purely optional, and the source of the report/concern is not revealed to the person in distress. The resources can include, but are not limited to, country-based support hotlines, online chat resources, and tips and suggestions.

3. In rare and extreme circumstances, the reviewer can conclude that the person in distress is in imminent danger, with their life in jeopardy without immediate intervention. If the reviewer makes that determination, we contact appropriate local resources (first responders, or others, depending on the locale) to promptly attempt to intervene and remove any potential risk of death.

Using this reactive approach, we were able to regularly provide support and resources to a substantial number of people. But, after further research on suicide cases, we realized that many posts indicative of suicidal ideation were not being reported, even when those posts had been seen by large number of people. And in a purely reactive solution, if the post is never reported, then Facebook is at a loss to help in any way. This limitation made us aware that there was still ample room for improvement and that there was more that we could do to support people in these difficult moments. This led us to consider alternatives that might provide better coverage for getting help to people at risk. The question we posed ourselves was the following: Could and/or should we do better by improving ways to detect at-risk individuals expressing suicidal ideation prior to human reports?

3.3 Queue Prioritization

The natural next step was to consider using Machine Learning. The following sections explain how we went about creating our current suicide prevention tools, which use ML to expand our ability to get timely help to people in need.[38] The decision to proactively intervene was something we did not take lightly, so anything proactive needed to be high-precision. Using ML, we started testing pattern recognition to identify posts signaling high likelihood of suicidal ideation. We first applied ML as a way to prioritize the reports that were reviewed by our specially trained reviewers in the Community Operations team[39] to ensure they reviewed the most urgent or likely to contain actionable information. Before this was implemented, we relied on recency for report prioritization. Our goal was to make sure the posts we could act on got to our human reviewers as fast as possible. Because our goal was tailored to Community Operations' evaluation criteria we trained our ML algorithms on posts that were reported by users and the actions taken by our Community Operations team were used as labels. We routed these reports to a unique review queue and found it worked well; reports from the new queue that ultimately required action started getting actioned remarkably sooner. But what if we could help more people that were currently not being

[38] https://code.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

[39] Among other tasks, Community Operations team reviews reported content on Facebook. For more information about this team, see our Hard Questions blogpost "Who Reviews Objectionable Content on Facebook — And Is the Company Doing Enough to Support Them? https://newsroom.fb.com/news/2018/07/hardquestionscontentreviewers/.

supported by our tools? What if we did not have to wait on human reporting at all?

3.4 Proactive Reporting

Once our prioritization was refined, we saw the opportunity to use those same types of classifiers to reach people whose posts might not be reported by friends or who we might be able to reach before a friend reports the post to us. We continued to use reports from Facebook users as the basis of our training data, focused around text analysis of the post and its associated comments. Through AI, we evaluate posts first when they are posted and again whenever a comment is added. We look at comments because a friend's reaction can help distinguish between a joke and post with serious intent. When a certain threshold is reached we then send a report to Community Operations who reviews the post and—if appropriate (and in the same way as if the report came from users)—sends resources to the person who posted the content, or reports it to first responders.[40] Right from the start of our early efforts, we found that these reports more accurately identified content that we were in a position to help with than user reports.

4 Why Machine Learning?

We decided to use Machine Learning because it provided the most nuance in determining the suicidal intent of a post. Our first attempts at classifying content used a simple list of words that could relate to ideation, but it quickly became clear we needed a more nuanced approach, and this led us to Machine Learning. For example, people will use a phrase like "If I hear that song one more time, I'm going to kill myself," and a word list containing the word kill, or the phrase "kill myself" is completely unable to discern between serious and sarcastic posts. We achieved a more nuanced understanding of what is a suicidal pattern and what is not by leveraging not only positive examples of suicidal ideation, but also—equally important—a small and precise set of negative examples[41]: the set of Facebook posts that people had flagged as potentially containing thoughts of suicide, but which the trained Community Operations reviewers determined did not demonstrate a person at risk of committing self-harm.[42] This set of negative examples contained a lot of the "I have so much homework I want to kill myself" type, which led to more precise training of the classifiers on accurate suicidal expressions. This enabled us to better distinguish between sarcastic and serious expressions of suicidal ideation, rendering our models more robust and accurate.

We went about training once again using our existing data set of Community Operation reports. We worked with outside experts in suicide prevention to select some of the most predictive features of suicidal expression. For example, the time of content

[40] Since these efforts began in 2017, Facebook has worked with first responders on over 1000 wellness checks based on reports received from our proactive detection methods. https://www.facebook.com/fbsafety/videos/1497015877002912/

[41] Machine Learning classifiers are trained with both positive examples (what the machine should identify) and negative examples (what the machine should not identify). In this way, the classifier learns to distinguish patterns between the two.

[42] https://newsroom.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/

posted was a feature we got from external experts. We also worked with internal data analysis to determine other features. Our Applied Machine Learning team helped us determine the best possible algorithms. Our first attempt at text classification was using a learning algorithm known as n-gram linear regression to find high scoring words in both posts and comments along with a random forest classifier that combines those classifiers with other features. Features like the time of day, content posted (photos, videos, plain text) or reactions were added into the random forest as a result of the external partnerships and internal data analysis. Ultimately, the more features the classifier could learn from the better, and the random forest determined the importance each feature had in evaluating the likelihood of any given post showing suicidal intent.

We also began to use an ensemble of classifiers that used different sets of features or that were trained on different data sets (different languages, text posts vs. videos) to find a wider array of content. During this process, we began to explore other algorithms, incorporating DeepText, a type of text classifier that can look at deeper meaning in a text by understanding similar words into our random forest classifiers. We actually found that using a combination of DeepText and linear regression in the random forest was the most effective.

5 Ethical Questions and the Balancing of Values

The evolution of Facebook's approach to suicide prevention was shaped by a series of ethical questions that we identified and addressed. In effect, the technical and product level decisions that were made during this process involved ethical balancing decisions between competing values and interests. This section elaborates on the ethical questions within our product development process, and how we reached the appropriate and justifiable balance between the competing values posed by those ethical questions.

5.1 Privacy Program

At Facebook, we consider diverse perspectives when building and reviewing our products, incorporating guidance from experts in areas like data protection and privacy law, security, content policy, engineering, product management, and public policy. This process, which involves a number of different teams (for example, legal, policy, privacy program, security, communications, and marketing) addresses potential privacy and other legal issues. It is important to note that this process goes beyond the assessment of legal compliance requirements, looking also into policy, ethical and societal implications of our products and services, while ensuring that our product applications bring value to people on our platform. The process starts at the product ideation phase and it goes until deployment, including—in some instances—post-deployment monitoring. During that process, we get wider input if necessary—this might include external policy briefings and consultation with industry and academic experts. It is within the product cross-functional review process that both privacy and ethical questions are discussed and settled upon. Discussion of ethical implications of our products and services are often a component of Facebook's privacy program. Our evaluation commonly goes beyond strictly privacy, data protection and other legal questions,

addressing product, and research use cases' decisions that raise competing values and interests for which there is no simple, straightforward, or pre-delineated solution. These ethical discussions also stress the need to create optimal outcomes in the wider interests of society, looking at the collective dimension of automation and algorithmic processes.[43] In the following, we'll describe and provide examples of concrete product decisions, within Facebook's approach to suicide prevention, that raised ethical questions and prompted the need to balance different values and interests.

5.2 Ethical Questions

5.2.1 Ethical Imperative?

The very first question we had to ask ourselves was why should we be doing this in the first place? In other words, why should we be deploying resources and building suicide prevention tools? Are we crossing lines and overstepping bounds, going beyond what we are supposed to do as a company?

We addressed this foundational question by considering a number of important factors.

First of all, as noted above, a number of suicide prevention institutions came to us and highlighted our unique position to help tackle this problem. From that moment, our answer, be it negative or positive, would inevitably underline an ethical positioning from our end. To answer that question and based on the research we conducted (as mentioned also above), we first recognized that we are, in effect, well positioned and resourced to build suicide prevention tools. By being a platform that people may use to express thoughts of suicide, and by connecting people to those who care about them, we have in place the social infrastructure necessary to use those connections in cases where people are expressing thoughts about suicide on Facebook. Secondly, this decision also stems from our mission to keep our community safe,[44] and is aligned with our overall focus, work, and investment on safety and well-being, which makes suicide prevention work an ethical imperative for Facebook.[45] This decision provided the ethical grounding and justification to start working in this space. From here, and as explained below, a number of more concrete ethical questions regarding particular and technical decisions on how to build and implement a tool for suicide prevention followed suit.

[43] "Current ethical debates about the consequences of automation generally focus on the rights of individuals. However, algorithmic processes—the major component of automated systems—exhibit a collective dimension first and foremost." In "Ethics and algorithmic processes for decision making and decision support" (AlgorithmWatch Working Paper No.2), Lorena Jaume-Palasí and Matthias Spielkamp.

[44] "Safe Community: As we build a global community, this is a moment of truth. Our success isn't just based on whether we can capture videos and share them with friends. It's about whether we're building a community that helps keep us safe—that prevents harm, helps during crises, and rebuilds afterwards… When someone is thinking of suicide or hurting themselves, we've built infrastructure to give their friends and community tools that could save their life," in https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/

[45] https://newsroom.fb.com/news/2017/12/hard-questions-is-spending-time-on-social-media-bad-for-us/

5.2.2 Privacy vs Efficacy

When building suicide prevention tools, one of the balances we need to attain is between efficacy and privacy. These interests may be at odds with each other, as going too far in efficacy—detecting suicidal ideation at all costs, with no regards to limits or boundaries—, could compromise or undermine privacy, i.e., the control over your own information and how it is used. The question we were faced with was the following: How can we deploy a suicide prevention system that is effective, and that protects people's privacy, i.e., that is not intrusive and respectful of people's privacy expectations?

For the sake of efficacy, there were a number of decisions that could have been made in order to optimize that goal. One of the possible technical options was to leverage all types of content that we have available to train the Machine Learning model and use in production. This would mean, for example, all posts regardless of privacy or group setting. Another technical option was to involve more actively the network of people and friends connected to the person that expressed suicidal ideation, notifying them proactively of such suicidal expressions and encouraging them to interact with the person in distress. Following these options would render the system more effective, but it would also undercut important privacy values and safeguards.

With that in mind, we deliberately decided not to use "only me" posts in training the suicide prevention classifier, as model inputs, and in deploying it afterwards in our platform. "Only me" posts are posts with the privacy setting selected to only the person posting the content: no one else besides the person uploading the post can see the content of the post. This means that the content needs to have been posted to friends, list of friends or more publicly, to be used in building the suicide prevention tool and using it in production. We also deliberately decided to exclude secret group posts because there are higher privacy expectations with this type of content. Content posted in the ambit of a secret group[46] were thus removed from the creation and deployment of the suicide prevention system. In a nutshell, and as a way to strike an ethical balance between privacy and efficacy, we decided not to look at "only me" posts or secret groups, but only at cases where a person is clearly reaching out to their friends.

Another important decision was not to highlight to anyone other than the authors themselves when their post has been flagged as expressing risk of suicide. An approach favoring efficacy would point towards proactively notifying friends about posts expressing suicidal ideation, encouraging them to engage with the person in distress that authored such post. But privacy considerations directed us to not issue any sort of proactive notifications and rely on our standard reporting mechanisms. Posts flagged as expressing suicidal ideation by our Machine Learning classifier will still trigger a recommendation on materials and helplines that the authors of the post, and only themselves, can leverage; but it will not alert in any way anyone else. Friends or other people in the audience of such posts can always, and on their own, report them as referring to suicide or self-injury.

Due to privacy considerations, we also focused primarily on the content of the posts rather than general risk factors tied to the user. In other words, we analyze the content of the posts without associating or collecting information on people uploading those posts.

[46] https://www.facebook.com/help/220336891328465?helpref=about_content

Our AI systems are trained to detect expressions of suicide without storing and connecting that information to specific people. The focus of our Machine Learning work is on the content expressed and not on the person expressing the content.

Last, but certainly not the least, we have been open and public about the deployment of this feature and the technical details involved in its building and rolling out. We provide explanations in our Help Center[47] and Safety Center[48] about the tools we have built for this purpose,[49] along with the Suicide and Self-Injury Resources. And we share resources in our Newsroom website with posts including "How Facebook AI Helps Suicide Prevention,"[50] "Getting Our Community Help in Real Time,"[51] and "Building a Safer Community with New Suicide Prevention Tools."[52]

The consistent theme to these decisions was to protect people's privacy while still trying to be as effective as possible, getting people help when we can. The privacy–efficacy alignment exercise will be an ongoing one so we can continue to strike the right balance as technology evolves.

5.2.3 Human in the Loop

Another question that was raised during the product development process revolved around the scope of human intervention and the extent to which we could rely on automation in the review of human generated reports. The tradeoff we faced in this particular context was the one between "full automation" and "human indispensability," or—as more popularly known—the question of the human in the loop. In our particular product building process, the questions boiled down to the following ones: Should all user reports go through human review, even those with low accuracy? Is there place for full automation, that is, delegation of the decision to not act on reports to the machine? Or to put it in other words: should we focus the resources, time, and energy of human review on reports that score high in accuracy, and auto-ignore those that score low?

Our internal analysis indicate that user reports on posts expressing suicidal thoughts tend to score low in accuracy and that most reactive reports are not actionable, that is, do not warrant specific actions from our Community Operations reviewers. Dismissing automatically low scoring reports would save resources and human review cost. Excluding such reports from manual review would allow our human reviewers to spend less time, attention and energy going through non-actionable reports, which only increases their fatigue and affects their productivity and motivation. Instead, and by fully automating the dismissal of these reports, human reviewers could focus on the reports that matter more, being more productive and strategic in the work they have been assigned to do.

Despite how appealing and apparently rational the decision to automatize low-scoring reports would be, namely in terms of economic savings and human

[47] https://www.facebook.com/help/1553737468262661?helpref=faq_content
[48] https://www.facebook.com/safety/wellbeing/suicideprevention
[49] "Why am I seeing messages about support tips and resources?" in https://www.facebook.com/help/382406778796431?helpref=search&sr=7&query=suicide
[50] https://newsroom.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/
[51] https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/
[52] https://newsroom.fb.com/news/2017/03/building-a-safer-community-with-new-suicide-prevention-tools/

resources optimization, there are important implications of removing the human from the loop that we need to take into consideration in the context of suicide prevention. The most important one is the cost of false negatives. If we automatically dismiss a report erroneously, categorizing it as non-actionable when in fact it is actionable, this could result in ignoring a person at risk. Despite the very low statistical probability, the cost of misclassifying a report and of doing that automatically with no human confirmation was deemed to be too high. In this light, we opted for the path of human indispensability, keeping the human in the loop. We thus chose to never automatically close reactive human generated reports, despite the fact that they are often far less actionable than proactive, machine driven ones. As such, all reports—even low-scoring, human-generated reactive ones—, need to go through human review. Guiding this decision was the severity of harm to people in our platform[53] that could derive from misclassifying a report without human intervention.[54]

5.2.4 Targeted and Effective vs Thorough and Ineffective (How Do We Set the Threshold?)

Once a ML model is created, there is an evaluation phase to assess the performance quality of the model, that is, to assess if the classifier is good enough to put it to use. In this case, the ML model generates a numeric prediction score for each post (estimate), and then applies a threshold to convert these scores into binary labels of 0 (low likelihood of suicidal risk) and 1(high likelihood of suicidal risk). By changing the score threshold, you can adjust how the ML model assigns these labels, and ultimately determines if that post is expressing thoughts of suicide.

As always with Machine Learning, there are tradeoffs between the recall (number of posts expressing an interest in suicide) and precision (the number of posts detected expressing an interest in suicide for compared to those that were not). If we wanted to ensure we caught every single post expressing suicidal intent then we would want to review every post put on Facebook, but of course that is impossible. ML is probabilistic in nature so it will never be possible to ensure 100% of accuracy in its use. So the question we had ahead of us was how can we target the relevant posts and allocate the strictly necessary resources for that, while being as thorough as we can? The question was about target-ness vs thoroughfulness.

Since we have decided we should find content expressing suicidal thoughts, we have to decide how many posts can be reasonably looked at in a day and the number of false positives we're willing to have our human reviewers look at. In ML terms, this is a question of how to set the threshold:

[53] See Singapore's Personal Data Protection Commission (PDPC) "Discussion Paper on Artificial Intelligence and Personal Data – Fostering Responsible Development and Adoption of AI" (published June 5th 2018) for a proposed decision matrix structure into 3 possible models (human in the loop; human over the loop; and human out of the loop). Such decision matrix takes into account the probability and severity of harm to consumers. https://www.pdpc.gov.sg/Resources/Discussion-Paper-on-AI-and-Personal-Data

[54] At a broader level, decisions regarding delegation to human review will be done on a product basis and will be dependent on the assessment of the type of data being processed, Machine Learning algorithm accuracy levels, and probability and severity of potential harms.

- If we lower the threshold, the more posts that will less likely be actionable will need to be reviewed by more people; this poses the risk of having a disproportionate number of human reviewers looking at non-concerning posts.
- If we raise the threshold, the more accurate will these posts be and the fewer people we will need to do the human review of the content; but this runs the risk of missing content that should have been flagged and reviewed.

In response to this challenge, our philosophy has been to maximize the use of human review available to us without falling beneath a certain threshold of accuracy. We have thus substantially increased our staffing in this area so that we can handle proactively reported posts of more people at risk. But we did this while maintaining high auto-report accuracy. Overall, we want to maintain a baseline accuracy to make sure that posts from people who need help are seen as fast as possible and are not lost among benign posts. Using human review allows us to run our classifiers at a lower precision but higher recall and ultimately help the most people. In this process, it's important for human reviewers that the content they're looking at is a mix of positives and negatives. Like ML algorithms, humans become less accurate when the content they're trying to identify is scarce.

Human reviewers always make the final decision on whether we send resources for reported content. Moreover, people can always report any posts, including ones expressing thoughts of suicide.

In addition, and when introducing a new classifier, we did not want to AB test it against our old classifiers. This was of particular concern during experimental design: if the new signal performed significantly worse we would miss out on helping people. Instead, we give each classifier the opportunity to report content, the newest classifier given the first opportunity, so we can establish a baseline for how accurate it is. We can remove a classifier when it is primarily reporting false positives because newer classifiers are catching the concerning content. This has ensured that when introducing new classifiers, people we could have helped won't get missed because we tested a faulty classifier.

5.2.5 Potential for Support vs Risk of Contagion (When Do We Leave Up or Take Down Posts/Videos?)

An important issue for us to think through was weighing the potential to provide support to the person expressing thoughts of suicide against the risk of contagion to those who are viewing the post or video. This question was particularly salient for Facebook Live,[55] where someone in distress may be sharing a video of themselves attempting suicide in real time. In this situation, the person contemplating or attempting suicide is in a very serious situation where harm to themselves may be imminent. When discussing this with experts, they pointed out that it could actually be helpful that this person is online and sharing their situation, instead of by themselves at home alone, as it enables other people to see the situation, reach out, and provide help. In this way, Facebook Live can become a

[55] Facebook Live lets people, public figures, and Pages share live video with their followers and friends on Facebook. For more info, see https://live.fb.com/about/.

lifeline to provide support to people in distress in the moment they need it the most. This simply wouldn't be possible had the person been at home alone without sharing it. The concern, however, is that viewing a suicide attempt or completion can be traumatic, and if left up to circulate provides the risk of contagion.

So how do we decide whether to leave up a post/video with a suicide attempt or to stop the Live stream or remove a post/video? We worked extensively with experts to provide guidelines around when a person may still be helped versus when it's likely a person can no longer be helped (i.e., person performs very specific behaviors that indicate a suicide attempt is in progress and they would no longer be responsive to help). If we determine that the person can no longer be helped and harm is imminent, we stop the Live video stream or will remove the post/video. At this point, the risk for trauma to the viewer is high, and we want to remove the post to mitigate the risk of contagion by it being viewed or shared. We do not take down the post/video as long as we think exposure may help. The hope is that someone will see the post or video and comment, reach out to the person directly, or flag it to Facebook to review so that we can provide resources.

6 Conclusion

Research shows that social support can help prevent suicide.[56] Facebook is in a unique position to connect people in distress[57] with resources that can help. We work with people and organizations around the world to develop support options[58] for people posting about suicide on Facebook, including reaching out to a friend, contacting help lines and reading tips about things they can do in that moment. Facebook also recently released suicide prevention support on Facebook Live[59] and introduced Artificial Intelligence to detect suicidal posts[60] even before they are reported.

The use of AI to help prevent suicide triggered a number of important and complex ethical questions that were tackled within Facebook's Privacy Program. Ethical questions do not lend themselves to easy or straightforward answers, or otherwise they would not be ethical in nature. In fact, these types of questions do not have pre-defined answers or solutions. These are not precise mathematical equations in search of a uniquely correct solution. Instead, these questions require us to balance and ponder between values, issues and interests in accordance with our mission and institutional practices.

In the particular case of Facebook's suicide prevention tools, we were driven by external conversations, by thoughtful research on how our role and position as a global

[56] Social support as a protective factor in suicide: Findings from two nationally representative samples Kleiman, Evan M. et al. *Journal of Affective Disorders*, Volume 150, Issue 2, 540–545 https://www.jad-journal.com/article/S0165-0327(13)00088-8/fulltext

[57] https://www.facebook.com/fbsafety/videos/1098099763561194/

[58] https://www.facebook.com/fbsafety/photos/a.197686146935898.42079.125459124158601/1041262189244952/

[59] https://newsroom.fb.com/news/2017/03/building-a-safer-community-with-new-suicide-prevention-tools/

[60] https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/

social network could help tackle this problem, and by the values of safety and well-being that we have prioritized and invested in.

Despite the fact that these ethical questions come without defined and predicted outcomes, companies need to put in place standardized and accountable processes to reach ethically justifiable and consistent answers. This is the case of Facebook's Privacy Program, which involved diverse perspectives early in the process to address ethical and privacy questions.

There is still work to be done. We continue to iterate with experts in the field to improve the resources we offer. We also integrate with partners like Crisis Text Line and Lifeline to provide online chat, which is becoming more prevalent over time.

This article showcases how profound philosophical and ethical questions can be mapped all the way to specific and concrete product decisions. It does so by providing a case study example of how such ethical review processes work in a real corporate setting with a concrete product powered by a specific technology: AI and suicide prevention. We thus listed the ethical questions that were discussed, the underlying competing values animating each of those questions, and the thought process and approach leading to their resolution. We encourage other industry actors to engage in this level of analysis and to increase the transparency of their own internal procedures, explaining—for products and services that raised important ethical questions—how they reached and justified their answers.