Artificial Intelligence-Based Suicide Prediction

Mason Marks*

ABSTRACT

Suicidal thoughts and behaviors are an international public health problem contributing to 800,000 annual deaths and up to 25 million nonfatal suicide attempts. In the United States, suicide rates have increased steadily for two decades, reaching 47,000 per year and surpassing annual motor vehicle deaths. This trend has prompted government agencies, healthcare systems, and multinational corporations to invest in artificial intelligence-based suicide prediction algorithms. This article describes these tools and the underexplored risks they pose to patients and consumers.

AI-based suicide prediction is developing along two separate tracks. In "medical suicide prediction," AI analyzes data from patient medical records. In "social suicide prediction," AI analyzes consumer behavior derived from social media, smartphone apps, and the Internet of Things (IoT). Because medical suicide prediction occurs within the context of healthcare, it is governed by the Health Information Portability and Accountability Act (HIPAA), which protects patient privacy; the Federal Common Rule, which protects the safety of human research subjects; and general principles of medical ethics. Medical suicide prediction tools are developed methodically in compliance with these regulations, and the methods of its developers are published in peer-reviewed academic journals. In contrast, social suicide prediction typically occurs outside the healthcare system where it is almost completely unregulated. Corporations maintain their suicide prediction methods as proprietary trade secrets. Despite this lack of transparency, social suicide predictions are deployed globally to affect people's lives every day. Yet little is known about their safety or effectiveness.

* Assistant Professor, Gonzaga University School of Law; Affiliated Fellow, Yale Law School Information Society Project; Doctoral Researcher, Leiden Law School Center for Law and Digital Technologies. Many thanks to Katherine Strandburg, Ann Bartow, Ari Waldman, Andrea Matwyshyn, Ido Kilovaty, Thomas Kadri, and Roger Ford for their helpful comments on an earlier draft of this article. Thank you to Abbe Gluck, Jack Balkin, Katherine Kraschel, Adam Pan, Phillip Yao, the Yale Solomon Center for Health Law & Policy, and the Yale Information Society Project for the opportunity to discuss this article at the "Law and Policy of AI, Robotics and Telemedicine in Health Care" conference at Yale Law School. Special thanks to Joel Reidenberg, the Center on Law and Information Policy at Fordham Law School, and the Innovation Center for Law and Technology at New York Law School for the opportunity to discuss the article at the Northeast Privacy Scholars Workshop.

Though AI-based suicide prediction has the potential to improve our understanding of suicide while saving lives, it raises many risks that have been underexplored. The risks include stigmatization of people with mental illness, the transfer of sensitive personal data to third-parties such as advertisers and data brokers, unnecessary involuntary confinement, violent confrontations with police, exacerbation of mental health conditions, and paradoxical increases in suicide risk.

INTRODUCTION 101

I. AI MAY BE USEFUL FOR SUICIDE PREDICTION 103

A. TRADITIONAL METHODS OF SUICIDE PREDICTION ARE INACCURATE 103

B. THE TWO PARALLEL TRACKS OF AI-BASED SUICIDE PREDICTION 104

1. MEDICAL SUICIDE PREDICTION 106

2. SOCIAL SUICIDE PREDICTION 107

II. AI-BASED SUICIDE PREDICTION POSES RISKS TO PATIENTS AND CONSUMERS 111

A. SAFETY RISKS 111

B. PRIVACY RISKS 116

C. AUTONOMY RISKS 117

1. CENSORSHIP 118

2. WARRANTLESS SEARCHES 120

III. CONCLUSION 120

INTRODUCTION

Suicide is a global public health concern. The World Health Organization (WHO) estimates it claims a life every 40 seconds and kills 800,000 per year.[1] Non-fatal suicide attempts may be 20 to 25 times more common. Both attempted and completed suicides take a large toll on families, communities, and healthcare systems, and they are on the rise.[2] In the U.S., suicide rates rose by 25% between 1999 and 2016, and half the states experienced a rise of over 30%.[3] Suicide is now the second leading cause of death in American teens, it kills more Americans each year than auto accidents or homicides, and it costs the U.S. economy over $69 billion dollars a year.[4]

To address the growing suicide problem, governments, healthcare systems, and corporations are developing artificial intelligence (AI) based suicide prediction tools. In theory, suicide can be prevented if it can be accurately predicted. Yet in practice, predicting suicide is challenging because it is a complex problem with many contributing factors. Traditional methods of prediction involve medical checklists and questionnaires that yield inaccurate results, often little better than a coin toss, or what would be expected due to chance.[5] AI shows promise for increasing the accuracy of suicide predictions.[6]

This article describes the current range of AI-based suicide prediction tools

---

1. World Health Organization, *Suicide Data*, http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/ (last visited Sep. 25, 2018).

2. American Foundation for Suicide Prevention, *Suicide Statistics*, https://afsp.org/about-suicide/suicide-statistics/ (last visited Oct. 14, 2018); World Health Organization, *Preventing Suicide—A Global Imperative*, http://apps.who.int/iris/bitstream/handle/10665/131056/9789241564779_eng.pdf;jsessionid=A1C3BA3BB3E15829DD187BD773E9A0CF?sequence=1 (last visited Oct. 14, 2018).

3. Centers for Disease Control and Prevention, *Suicide Rising Across the US*, https://www.cdc.gov/vitalsigns/suicide/ (last visited Aug. 21, 2018) (reporting that since 1999, half the U.S. states experienced an increase in suicide rates of over 30%); Sabrina Tavernise, *U.S. Suicide Rate Surges to a 30-Year High*, N.Y. TIMES (Apr. 22, 2016), https://www.nytimes.com/2016/04/22/health/us-suicide-rate-surges-to-a-30-year-high.html.

4. Alicia Vanorman and Beth Jarosz, *Suicide Replaces Homicide as Second-Leading Cause of Death Among U.S. Teenagers*, Population Reference Bureau (Jun. 9, 2016), https://www.prb.org/suicide-replaces-homicide-second-leading-cause-death-among-us-teens/; National Highway Traffic Safety Administration, *USDOT Releases 2016 Fatal Traffic Crash Data*, https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data (reporting that in 2016, nearly 45,000 Americans died by suicide. By comparison, 37,461 Americans were killed in auto accidents); Federal Bureau of Investigation, *Murder*, https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/murder (last visited Sep. 25, 2018) (reporting that in 2016, there were 17,250 U.S. homicides); Margot Sanger-Katz, *Gun Deaths Are Mostly Suicides*, N.Y. TIMES (Oct. 9, 2015), https://www.nytimes.com/2015/10/09/upshot/gun-deaths-are-mostly-suicides.html.

5. Colin G. Walsh et al., *Predicting Risk of Suicide Attempts Over Time Through Machine Learning*, 5 CLINICAL PSYCHOLOGICAL SCI. 1, 2 (2017).

6. *Id.*

and their underexplored legal, ethical, and public health implications. Healthcare systems including Vanderbilt University Medical Center, government agencies such as the U.S. Department of Veterans Affairs (VA), and private companies including Facebook are developing AI-based suicide prediction tools.[7] Though these tools have the potential to locate people at high risk for suicide, permitting intervention and possibly prevention, they also potentially violate people's privacy, marginalize vulnerable populations, stigmatize and traumatize people with disabilities, inaccurately categorize people as suicidal or non-suicidal, promote unnecessary civil commitments (and in some parts of the world incarceration), exacerbate mental health conditions, and paradoxically increase the risk of suicide. These risks have received little or no attention in the media and academic literature. As AI-based suicide prediction tools become more widespread, it is important to determine whether they are helping to prevent mental illness and suicide or if they are contributing to these problems.

The article consists of two parts. Part I explains why traditional methods of suicide prediction are inaccurate and how AI-based tools may improve upon their accuracy. These tools fall into two general categories: medical and social suicide prediction, which use different methods and draw from different data sets. Part I also describes how these two categories are governed by different laws leaving medical suicide prediction heavily regulated and social suicide prediction almost completely unregulated.

Part II describes the individual and societal risks of AI-based suicide prediction and how they may disproportionately impact vulnerable populations. The risks are divided into privacy, safety, and autonomy harms. Part II also explains how suicide prediction is analogous to predictive policing and suffers from similar shortcomings and misconceptions. This analogy is particularly strong because in some countries, attempted suicide is a criminal offense, and AI-based suicide prediction could result in criminal penalties including fines and imprisonment.

The article concludes with preliminary recommendations for minimizing the risks associated with AI-based suicide prediction tools.

7. Martin Kaste, *Facebook Increasingly Reliant on A.I. To Predict Suicide Risk*, ALL THINGS CONSIDERED (Nov. 17, 2018), https://www.npr.org/2018/11/17/668408122/facebook-increasingly-reliant-on-a-i-to-predict-suicide-risk.

I. AI MAY BE USEFUL FOR SUICIDE PREDICTION

A. Traditional Methods of Suicide Prediction Are Inaccurate

Traditionally, doctors and therapists predicted suicide by administering written questionnaires to patients. The answers were converted into scales intended to reflect suicide risk. Typical examples include the Suicide Intent Scale, the Scale for Suicidal Ideation, and the Beck Hopelessness Scale. However, their predictive abilities are unimpressive: "Recent meta-analyses of hundreds of studies from the past 50 years indicate that the ability to predict future suicide attempts has always been at near chance levels."[8] According to one large study: "All of the scales and tools reviewed here had poor predictive value."[9]

Suicide is difficult to predict because it is a complex problem with many risks and contributing factors.[10] There is no single risk factor that reliably predicts suicide. Though there is a clear association between suicide attempts and some variables such as a history of depression or a substance use disorder, most people with these conditions do not attempt or die by suicide.[11] Other potential suicide risk factors include substance use disorders, anxiety disorders, bipolar disorder, eating disorders, unemployment, a family history of suicide, having been released recently from a psychiatric hospital, "belonging to a sexual minority" group, "infection with the brain-tropic parasite *Toxoplasma gondii*," and "childhood physical, sexual, or emotional abuse."[12] Because the risk factors are so numerous and diverse, it is difficult to account for them all in a single predictive model.

Accurate suicide prediction is also hindered by the fact that suicide is relatively rare.[13] Though on a national and global scale, the number of people who die by suicide is by no means trivial, only a very small percentage of people under psychiatric care attempt suicide.[14] According to estimates by the U.S. Substance Abuse and Mental Health Service Administration, 9.8 million

8. Walsh, *supra* note 5.

9. Melissa K.Y. Chan et al., *Predicting Suicide Following Self-Harm: Systematic Review of Risk Factors and Risk Scales*, 209 BRITISH J. PSYCHIATRY 277, 279 (2016).

10. Gustavo Turecki and Brent A. David, *Suicide and Suicidal Behavior*, 387 LANCET 1227 (2016) (reporting genetic, developmental, and social risk factors for suicide).

11. *See, e.g., id.* at 6.

12. *Id.* at *5.*

13. *See* Steffan Davies et al., *Depression, Suicide, and the National Service Framework – Suicide is Rare and the Only Worthwhile Strategy is to Target People at High Risk*, 322 BMJ 1500 (2001).

14. Roger Mulder et al., *The Futility of Risk Prediction in Psychiatry*, 209 BRITISH J. PSYCHIATRY 271, (2016).

American adults seriously contemplated suicide in 2015.[15] However, only 2.7 million formulated concrete suicide plans and about 1.4 million made suicide attempts.[16] These statistics demonstrate that even though suicidal thoughts are a risk factor for suicide, most people who have suicidal thoughts do not attempt suicide.[17] The same can be said for major depressive disorder, which is estimated to affect over 16 million American adults.[18]

Suicide prediction is also challenging because discussing suicide is taboo. Patients with suicidal thoughts may be afraid to discuss it with friends, family, and healthcare providers out of fear they might be judged, stigmatized, or hospitalized and medicated against their will.[19] Even if people did disclose suicidal thoughts with healthcare providers more often, such reporting may not be the best predictor of an impending suicide because "the vast majority of individuals who express suicidal ideation never go on to attempt it."[20] Certain subpopulations may share cultural values that make discussion of suicidal thoughts more challenging. For example, the U.S. military's culture of promoting mental toughness, self-sacrifice, and the control and suppression of emotions can serve as an obstacle to open discussion of emotionally charged issues like suicide.[21] Service members may feel obligated to suppress their feelings and "shake it off" when facing feelings of despair.[22]

The next section explains how AI may increase our ability to identify people at risk for suicide and describes the two general tracks of AI-based suicide prediction.

B. The Two Parallel Tracks of AI-based Suicide Prediction

AI may overcome many limitations of traditional suicide screening tools and increase the accuracy of predictions. AI-based suicide prediction tools can be divided into two broad categories: The first category involves analysis of patient medical records. It is performed by doctors, public health researchers,

15. *9.8 Million American Adults had Serious Thoughts of Suicide in 2015*, SUBSTANCE ABUSE MENTAL HEALTH SERVICES ADMIN. (Sep. 15, 2016), https://www.samhsa.gov/newsroom/press-announcements/201609150100.

16. *Id.*

17. *Id.*

18. *Major Depression*, NAT'L INST. MENTAL HEALTH, https://www.nimh.nih.gov/health/statistics/major-depression.shtml#part_155028 (last visited Jan. 2019).

19. Lindsay Sheehan et al., *The Specificity of Public Stigma: A Comparison of Suicide and Depression-Related Stigma*, 256 PSYCHIATRY RESEARCH 40 (2017).

20. Chris Poulin et al., *Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes*, 9 PLOS ONE e85733 (2014).

21. Chris Poulin et al., *Predicting Military and Veteran Suicide Risk: Cultural Aspects*, 9 PLOS ONE 1 (2014).

22. *Id.*

government agencies, hospitals, and healthcare systems. I will refer to this category as "medical suicide prediction" because it is based on medical records and is usually conducted within the healthcare system; The second category involves the analysis of consumer behavior and social interaction derived from retail purchases, smart phone apps, social media, and other activities outside of healthcare. I will refer to this category as "social suicide prediction" because it is based on data derived from people's interactions with each other mediated by technology.

In the U.S., medical and social suicide prediction are subject to different laws. For example, medical suicide prediction is governed by the Health Information Portability and Accountability Act (HIPAA), which protects patient privacy and imposes civil and criminal penalties on covered entities when patient records are breached. Medical suicide prediction research is subject to the Federal Common Rule, which safeguards human research subjects. All research must also comply with general principles of medical ethics and be approved by hospital and university institutional review boards (IRBs).

In contrast, social suicide prediction is usually subject to none of these requirements because it is conducted primarily outside the healthcare system. Because it involves making predictions about consumers, it is governed by agencies that protect consumers and regulate interstate commerce and communication such as the Federal Trade Commission (FTC), the Federal Communications Commission (FCC), the Food and Drug Administration (FDA), and the laws these agencies enforce. At least so far, the agencies have taken little or no interest in suicide predictions and their associated risks. In contrast, regulatory agencies in the UK have taken an interest in how social media platforms predict and prevent suicide and other forms of self-harm.[23]

A few groups engage in both medical and social suicide prediction. For instance, the VA has analyzed both veterans' medical records and their social media activity.[24] When doctors and hospitals make social suicide predictions, those predictions are subject to the laws that typically cover medical suicide predictions. However, most groups conduct only one type of suicide prediction, and they can generally be divided into groups that reside within the healthcare system (medical suicide predictors) and those that do not (social suicide predictors). Though AI-based suicide prediction can be divided into these two

23. *See, e.g.*, Mason Marks, *Censoring Self-Harm on Facebook Might Do More Harm Than Good*, MOTHERBOARD (Mar. 1, 2019), https://www.vice.com/en_us/article/d3m5vj/censoring-self-harm-on-facebook-might-do-more-harm-than-good.

24. Chris Poulin & Gregory Peterson, *Mobile and Social Networking Technology Monitors Big Data from Messages to Detect Suicide Risk in Military Veterans*, ELSEVIER (Nov. 11, 2015), https://www.elsevier.com/connect/artificial-intelligence-app-combats-suicide-in-veterans.

categories, there is considerable variation within each category with respect to the populations studied, the data collected, and the methods used.

The following sections describe the activities of the most prominent medical and social suicide predictors.

1. Medical Suicide Prediction

Medical suicide prediction uses AI to scan and analyze medical records. It is most often performed by academic medical centers, hospitals, and government agencies such as the VA.

In one of the largest studies to date, published in 2018, Simon et al. analyzed the anonymized records of nearly 3 million patients across seven health systems in multiple states.[25] The records included data from over 10 million mental health specialty visits and nearly 10 million primary care visits.[26] The authors report that their method of combining large volumes of medical records with data from standard mental health questionnaires outperformed previous methods using medical records alone.[27]

In 2017, a study of patient data from within a single healthcare system was published by Colin Walsh et al.[28] It analyzed the records of 5,167 adults treated at Vanderbilt University Medical Center.[29] The authors report the accuracy of their suicide prediction models in terms of accuracy under the curve (AUC) where an AUC of 0.5 represents "accuracy no better than chance" and an AUC of 1.0 represents perfect accuracy.[30] Remember that traditional methods of suicide prediction may be little more accurate than a coin flip (a probability of about 50% or 0.50). For patients attempting suicide for the first time, Walsh reported AUC values ranging from 0.82 "at 7 days prior to suicide attempts" to 0.75 "at 720 days prior to suicide attempts."[31]

A smaller study published in 2014 by Poulin et al. analyzed the clinical records of 100 veterans who died by suicide in 2009.[32] The study identified words and word pairs in clinical notes that were associated with suicide.[33] Predictive models based on single words, such as "agitation" and "analgesia,"

25. Gregory E. Simon et al., *Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records*, 175 AMER. J. PSYCHIATRY 951, 953 (2018).
26. *Id.*
27. *Id.* at 958.
28. Walsh, *supra* note 5.
29. *Id.*
30. *Id.* at 3.
31. *Id.* at 7.
32. Poulin & Peterson, *supra* note 24, at 2.
33. *Id.* at 3.

had an average predictive accuracy of 59%.[34] The predictive accuracy of word pairs ranged from 52% - 69%.[35]

The studies by Simon, Walsh, and Poulin illustrate the potential of AI to improve the accuracy of suicide predictions. However, their prediction models still have limitations. For instance, though they may represent an improvement over traditional methods, current medical suicide prediction models still produce a significant number of false positives and false negatives. As a result, they are primarily used for research and not to guide clinical decision-making.

Because medical suicide prediction tools are developed within the healthcare system, they are subject to the norms and regulations of that system. Before a medical suicide prediction study can commence, its protocols must be reviewed and approved by IRBs that aim to ensure compliance with state, federal, and institutional safety and ethical standards. In contrast, the tools described in the following section lack these safeguards, yet they are used every day to impact real people's lives.

2. Social Suicide Prediction

The medical suicide prediction studies listed above utilized medical records to train their AI prediction models. This method can be contrasted with the approach of tech companies such as Facebook, Crisis Text Line, and Objective Zero. Unlike healthcare providers and medical researchers, these companies lack access to patient records. Instead, they have access to large data sets derived from the behavior of users. When consumers browse the Internet, shop online, use ride-sharing apps like Uber or Lyft, stream music and video, or post on social media, they leave behind trails of digital traces that reflects where they have been and what they have done. Companies collect these digital traces and analyze them with AI to reveal people's sensitive health information.[36] The goal of social suicide prediction is to infer suicide risk from people's digital traces. The most prominent example is Facebook's system. Other examples include Crisis Text Line, and Objective Zero.

Since Facebook introduced its live-streaming service "Facebook Live" in early 2016, dozens of users have broadcast suicide attempts in real-time on the platform.[37] On February 16, 2017, Facebook CEO Mark Zuckerberg announced

34. *Id.*

35. *Id.*

36. *See* Mason Marks, *Emergent Medical Data*, PETRIE-FLOM CTR.: BILL OF HEALTH (Oct. 11, 2017), blog.petrieflom.law.harvard.edu/2017/10/11/emergent-medical-data/.

37. *See* Nicolas Vega, *Facebook: We Can't Stop All Live-stream Suicides*, N.Y. POST (Oct. 25, 2017), https://nypost.com/2017/10/25/facebook-we-cant-stop-all-live-stream-suicides/; Jessica Guynn, *Facebook Live is Scene of Another Suicide; Police Say 'I Hope This isn't a Trend'*, USA

the company was developing AI to analyze and flag user-generated content for review by its community managers.[38] In this announcement, Zuckerberg mentioned suicide prediction and prevention as one of his priorities. On March 1, 2017, Facebook announced its application of AI to identify suicidal intent in user-generated content.[39] According to a company spokesperson, machine learning algorithms scan users' posts, and comments made in response to those posts, for cues that reflect elevated suicide risk.[40] In a Facebook promotional video released on November 26, 2017, the Chautauqua County Sheriff's Department in Upstate New York praises Facebook for alerting it to a potential suicide, which enabled officers to intervene.[41] The following day, Facebook announced its AI-based suicide prediction program had initiated over 100 such "wellness checks," which are often referred to as welfare checks by the law enforcement community. In that announcement, Facebook said it would expand its suicide prediction program globally in "most of the countries in which it operates, with the exception of those in the European Union (EU)."[42]

On April 2, 2018, Zuckerberg revealed that Facebook's AI scans the contents of users' private messages, which suggests that both public and private user-generated content may be scanned for signs of suicidal intent.[43] On September 10, 2018, Facebook provided additional details about its suicide prediction algorithms: Using an AI tool called random forests, Facebook analyzes user-generated content and assigns a risk-rating to words, word pairs, and phrases in each post. Hypothetical examples provided by the company include "sadness," "much sadness," and "so much sadness." This method is like the approaches used by Walsh and Poulin. However, in Facebook's case, the words and phrases are derived from social media content instead of medical

---

TODAY (Apr. 26, 2017), https://www.usatoday.com/story/tech/news/2017/04/26/facebook-live-another-suicide/100941914/.

38. Diana Kwon, *Can Facebook's Machine-Learning Algorithms Accurately Predict Suicide*, SCI. AM. (Mar. 8, 2017), https://www.scientificamerican.com/article/can-facebooks-machine-learning-algorithms-accurately-predict-suicide/; Mark Zuckerberg, *Building Global Community*, FACEBOOK (Feb. 16, 2017), https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634.

39. Kwon, *supra* note 38; Vanessa Callison-Burch, *Building a Safer Community with New Suicide Prevention Tools*, FACEBOOK (Mar. 1, 2017), https://newsroom.fb.com/news/2017/03/building-a-safer-community-with-new-suicide-prevention-tools/.

40. Kwon, *supra* note 38.

41. FACEBOOK SAFETY, https://www.facebook.com/fbsafety/videos/1497015877002912/.

42. Hayley Tsukayama, *Facebook is Using AI to Try to Prevent Suicide*, WASH. POST (Nov. 27, 2017), https://www.washingtonpost.com/news/the-switch/wp/2017/11/27/facebook-is-using-ai-to-try-to-prevent-suicide/.

43. Sarah Frier, *Facebook Scans the Photos and Links You Send on Messenger*, BLOOMBERG (Apr. 4, 2018), https://www.bloomberg.com/news/articles/2018-04-04/facebook-scans-what-you-send-to-other-people-on-messenger-app.

records.

Unlike medical suicide prediction, which is mostly experimental, requires approval from IRBs, and results in peer reviewed studies in academic journals, Facebook's suicide prediction program is not subject to independent ethics review, and its methods and results are not published or otherwise made public. This lack of accountability and transparency raises safety concerns that are discussed in Part II. Instead of consulting an IRB, Facebook sometimes utilizes an internal ethics board.[44] However, unlike IRB approval at a hospital or university, which is mandatory, review of Facebook's projects by its ethics board occurs at the company's discretion.[45]

Facebook's lack of transparency and accountability is concerning because the company has a history of monitoring people's emotional states and experimenting on users without their knowledge or consent.[46] Since the company's wellness checks were made public in late 2017, Facebook has expanded its suicide prediction program internationally and conducted at least 3,500 wellness checks in the U.S. and abroad.[47]

Many questions about Facebook's social suicide prediction program remain unanswered. For example, on what data were its algorithms trained? Facebook provides only vague answers. According to an article written by its software engineers: "To start, we worked with experts to identify specific keywords or phrases known to be associated with suicide."[48] However, Facebook quickly learned this approach resulted in too many false positives, picking up benign phrases such as "Ugh, I have so much homework I just wanna kill myself," which is meant to express frustration rather than suicidal intent.[49]

Facebook then implemented an AI-based approach using machine learning. According to its engineers: "We were able to use posts previously reported to Facebook by friends and family, along with the decisions made by our trained reviewers (based on our Community Standards), as our training data set."[50] This quote reveals a serious limitation of Facebook's AI training method. Because the company lacks access to medical records, it cannot train its AI using data from

44. Molly Jackman and Lauri Kanerva, *Evolving the IRB: Building Robust Review for Industry Research*, 72 WASH. & LEE L. REV. ONLINE 422 (2017).

45. *Id.*

46. *See* Robinson Meyer, *Everything We Know About Facebook's Secret Mood Manipulation Experiment*, ATLANTIC (Jun. 28, 2014), https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/.

47. Kaste, *supra* note 7.

48. Dan Muriello et al., *Under the Hood: Suicide Prevention Tools Powered by AI*, Facebook Code (Feb. 21, 2018), https://code.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/.

49. *Id.*

50. *Id.*

actual suicides. Instead, it appears to use the reports of concerned Facebook users and the subsequent actions of its content moderators as a proxy for suicide risk. Facebook's approach has severe limitations because instead of accurately predicting suicidal thoughts and behaviors, Facebook's AI may merely be predicting what its users and content moderators perceive to be suicide risk.

In an e-mail interview, I asked Facebook's Emily Cain whether the company retains information about the outcomes of wellness checks. She said: "Most of the time we do not know the outcome of those wellness checks because first responders usually keep that information confidential. On occasion, first responders will respond to Facebook's escalation to share the outcome of the intervention." Thus, Facebook may receive some suicide data from emergency responders following wellness checks. However, it is unknown whether the company feeds this data back into the system to improve its suicide predictions.

A lack of real-world suicide data would significantly reduce the accuracy of Facebook's predictions. I asked whether the company conducts experiments to test their accuracy. According to Cain, "We audit and perform quality checks to ensure that we're moderating content for suicide and self-injury appropriately (taking down violating content, checkpointing people who appear to be in crisis, escalating imminent issues) the same way we do across all content on Facebook."[51] However, she declined to provide further information regarding these processes.

When asked what training and certification Facebook's content moderators have and what criteria they use to decide when police should be contacted, Cain responded:

> Our Community Operations team includes thousands of people around the world who review reports about content on Facebook. The team includes a dedicated group of specialists who have specific training in suicide and self-harm . . . Where we have signals of potential imminent risk or harm, a specialized team conducts an additional review to determine if we should help refer the individual for a wellness check. Those teams are trained to engage directly with first responders to assist them in locating the person to conduct a wellness check. This team has experience in safety, law enforcement response, or crisis response with backgrounds in domestic and federal U.S. law enforcement, rape and suicide hotlines, Center for Missing or Sexually Exploited Children, Social Services, international law enforcement as well as domestic and international crisis and

51. E-mail interview between Mason Marks and Facebook spokeswoman Emily Cain.

intervention centers.

It may seem reassuring that Facebook's community operations team includes people with experience working in crisis intervention. However, without more information about their credentials and how they make decisions to escalate cases to police, it is difficult to evaluate the program.

II. AI-BASED SUICIDE PREDICTION POSES RISKS TO PATIENTS AND CONSUMERS

AI holds promise for improving suicide predictions. However, it exposes people to a variety of dangers, which can be divided into safety, privacy, and autonomy risks: Safety risks include false negatives that may leave suicidal individuals without assistance, false positives that can cause biased treatment by physicians, unexpected and unwarranted visits from police that may escalate to violent confrontations, and involuntary medical treatment; privacy risks include the leak of sensitive information through security breaches, and the transfer or sale of personal data to third parties such as data brokers and advertisers, which can lead to stigmatization, exploitation, and discrimination; and, autonomy risks include censorship, unnecessary confinement or civil commitment, and in countries where suicide attempts are illegal, criminal penalties including fines and incarceration. The following sections describe these risks in greater detail. There is significant overlap between the risks of medical and social suicide prediction. For this reason, the risks of both categories are discussed together. However, because medical suicide prediction is governed by health laws and regulations, people subjected to it are provided greater protection than those who are subjected to social suicide prediction.

A. Safety Risks

The safety risks of AI-based suicide predictions stem from their inaccuracy and the limited effectiveness of interventions that are triggered by predictions. Despite purported improvements over traditional prediction methods, AI-based predictions produce many false positives and false negatives.[52] Both types of misclassification can affect people's safety. The risks associated with false negatives are easiest to understand. If suicide predictions are less than 100 percent accurate, they will inevitably fail to identify some suicidal people. Those individuals might not receive needed assistance and may harm or kill themselves.

By comparison, the safety risks of false positives are more complex. They stem from stigmatization and the treatment interventions that result from being

52. Mulder, *supra* note 14.

labeled high-risk for suicide. People placed in this category may be treated differently by physicians in ways that endanger their health and safety. Dr. Greg Simon likens false positives from suicide prediction algorithms to false positives from vehicle blind spot warning systems. If a blind spot warning system issues false positives, the driver can act as though they are true positives, postpone switching lanes, and little harm is done. In the worst-case scenario, he might miss his off-ramp and have to double back. This analogy may hold true in limited cases. For instance, if the result of a false positive is a non-invasive, soft-touch intervention, such as providing referrals to counseling centers, the harm to a patient or consumer may be minimal. However, for the most part, Simon's analogy is a poor fit for suicide predictions. If suicide screening tools produce false positives, there may be long-lasting and potentially fatal adverse effects.

According to an article in the British Journal of Psychiatry: "The most obvious harm is that patients labelled 'high risk' may receive needlessly more restrictive treatments."[53] For example, patients might be taken off certain medications due to the perceived suicide risk even if the medications are helpful. One current example involves opioids. Despite the ongoing U.S. opioid crisis, opioids remain an appropriate treatment for many patients.[54] However, in the context of the crisis, physicians are increasingly reluctant to prescribe opioids, and if an algorithm labels a patient high risk for suicide, doctors might respond by withholding access to opioids due to the perceived risk of overdose.[55] Patients undergoing surgery may receive inadequate post-operative pain control, and patients prescribed opioids for chronic pain may be abruptly tapered off them. Thus, patients could unnecessarily be forced to endure pain and its complications due to inaccurate suicide predictions.[56]

Due to false positives, patients might be hospitalized against their will, and a diagnosis of suicidal thoughts would become part of their permanent medical

53. *Id.*

54. *See* Marilyn Serafini, *The Physicians' Quandary with Opioids: Pain versus Addiction,* NEJM CATALYST (Apr. 26, 2018), https://catalyst.nejm.org/quandary-opioids-chronic-pain-addiction/.

55. *See* Juliann Garey, *When Doctors Discriminate,* N.Y. TIMES (Aug. 10, 2013), https://www.nytimes.com/2013/08/11/opinion/sunday/when-doctors-discriminate.html (reporting physician bias and refusal to prescribe pain medication following disclosure of bipolar disorder diagnosis).

56. Withdrawing adequate pain control may be inappropriate even if suicide predictions are accurate because many suicides have been blamed on physicians' tapering or withholding opioids resulting intractable pain. *See, e.g.*, Thomas Kline, *#OpioidCrisis Pain Related Suicides Associated with Forced Tapers*, MEDIUM (May 11, 2018), https://medium.com/@ThomasKlineMD/opioidcrisis-pain-related-suicides-associated-with-forced-tapers-c68c79ecf84d; Elizabeth Llorente, *As Doctors Taper or End Opioid Prescriptions, Many Patients Driven to Despair, Suicide,* FOX NEWS (Dec. 10, 2018), https://www.foxnews.com/health/as-opioids-become-taboo-doctors-taper-down-or-abandon-pain-patients-driving-many-to-suicide.

records. Healthcare providers may find it difficult to ignore the results of AI-based suicide predictions even when they disagree with the predictions and suspect they might be false positives. Similar concerns have been raised in the context of the justice system, where judges use opaque, proprietary algorithms in bail and sentencing hearings to predict who is likely to recidivate.[57] Though sentencing decisions are ultimately the responsibility of judges, they may be influenced by algorithmic assessments.

In the healthcare setting, doctors may be incentivized to follow AI-based suicide predictions because overriding a prediction could expose them to medical malpractice liability if they don't hospitalize patients who subsequently attempt or complete suicide.

Involuntary hospitalization and forced medication are not without risks. Though they can prevent suicide in the short term, unnecessary confinement and treatment may paradoxically increase suicide risk because the experience can be traumatic and dehumanizing.[58] People are at increased risk for suicide shortly after being admitted to hospitals and shortly after being released.[59] Moreover, it is well documented that numerous psychiatric medications are associated with transient increases in suicide risk. These risks may be exacerbated when people lack access to mental health resources and social support outside the hospital. There is also a risk that doctors will treat patients categorized as high-risk differently than other patients. Physicians are sometimes biased against patients with mental illnesses, substance use disorders, and histories of suicidal thoughts.[60] As a result, a false positive placed into a patient's record may affect how physicians treat the patient in the future resulting in sub-optimal care.

Healthcare providers, social media platforms, and police may over rely on suicide predictions. According to one metanalysis on suicide risk assessment: "[A]n over-reliance on the identification of risk factors in clinical practice, is, in our view, potentially dangerous and may provide a false reassurance for clinicians and managers."[61] The authors emphasize that clinicians should draw a distinction between risk assessment and prediction: "The idea of risk assessment

57. *See* Jason Tashea, *Courts are Using AI to Sentence Criminals. That Must Stop Now,* WIRED (Apr. 17, 2017), https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now.

58. *See* B. Olofsson and L. Jacobsson, *A plea for respect: involuntary hospitalized psychiatric patients' narratives about being subjected to coercion,* 8 PSYCHIATRIC MENTAL HEALTH NURSING 357 (2001); *see* also Gail C. Eisenberg, *Involuntary Commitment and the Treatment Process: A Clinical Perspective,* 44 BULLETIN AMER. ACAD. PSYCHIATRY L. (1980).

59. *See* Ping Qin and Merete Nordentoft, *Suicide Risk in Relation to Psychiatric Hospitalization,* 62 ARCH. GEN. PSYHICATRY 427 (2005).

60. Stephanie Knaak, *Mental illness-related stigma in healthcare,* 30 HEALTHCARE MANAGEMENT FORUM 111 (2017).

61. Chan et al., *supra* note 9.

as risk prediction is a fallacy and should be recognized as such. We are simply unable to say with any certainty who will and will not go on to have poor outcomes. People who self-harm often have complex and difficult life circumstances, and clearly need to be assessed- but we need to move away from assessment models that prioritise risks at the expense of needs."[62] Thus, the authors appear to advocate for a soft-touch approach to suicide prevention in which risk assessments lead to more thorough, individualized evaluations instead of firm-hand suicide interventions such as sending police to people's homes.

Firm-hand interventions including wellness checks could have unexpected consequences such as the exacerbation of symptoms and involuntary hospitalization. Police response may further escalate already tense situations. There are numerous reports of people being shot by police after they arrive to investigate erratic behavior or a threat of suicide. In some cases, it is believed suicidal individuals provoke police with the goal of being shot, which is termed "suicide by cop." In other cases, the reasons for police shootings are less clear.

On June 14, 2014, Jason Harrison's mother called Dallas police requesting their help transporting him to a hospital for psychiatric care.[63] Harrison was 38 years old and had been diagnosed with schizophrenia and bipolar disorder.[64] When police arrived, Harrison stood in the doorway holding a small screwdriver.[65] Despite carrying less-than-lethal weapons such as Tasers and pepper spray, two officers drew their firearms and shot and killed Harrison.[66]

On March 9, 2015, an Atlanta-area police officer shot and killed 27 year old Air Force veteran Anthony Hill.[67] According to Hill's family, he was experiencing a non-violent episode resulting from trauma endured while on active duty in Afghanistan.[68] Hill had previously been treated for bipolar disorder.[69] On the night of his death, police responded to reports that he had jumped from a second story balcony and was behaving erratically on the grounds

62. *Id.*

63. Curtis Skinner, *Family of Jason Harrison, Mentally Ill Man Killed by Dallas Police, Release Graphic Video,* Huffington Post (Mar. 17, 2015), https://www.huffingtonpost.com/2015/03/17/jason-harrison-shooting-v_n_6887242.html.

64. *Id.*

65. *Id.*

66. *Id.* (showing officer with Taser and pepper spray holstered on his utility belt)

67. Associated Press, *Atlanta-area police officer charged with felony murder for shooting of Anthony Hill,* GUARDIAN (Jan. 22, 2016), https://www.theguardian.com/us-news/2016/jan/22/anthony-hill-shooting-atlanta-georgia-police-felony-murder-charge-robert-olsen.

68. Yanan Wang, *Georgia Police Officer Indicted for Murder in Shooting of Unarmed, Naked Black Veteran,* WASH. POST (Jan. 22, 2016), https://www.washingtonpost.com/news/morning-mix/wp/2016/01/22/georgia-police-officer-indicted-for-murder-in-shooting-of-unarmed-naked-black-veteran/?utm_term=.a05bbf170323#comments.

69. *Id.*

of an apartment complex.[70] When police arrived, Hill approached an officer while naked and unarmed.[71] Though the officer carried a Taser, he drew his firearm before shooting and killing Hill.[72]

Three examples from 2018 illustrate the dangers of relying on third-party reports from social media to initiate wellness checks. In each case, police may have responded with aggression out of proportion to the risk posed to them. On January 20, 2018, high school student John Albers was shot and killed by police responding to a 911 call claiming he threatened to kill himself during a video chat session on Apple's Facetime.[73] The dispatcher informed police that Albers was alone and in the basement of his family's home.[74] According to police, as they approached the home, a garage door opened, and a vehicle emerged and moved towards one officer.[75] The officer fired 13 shots into the family minivan killing Albers.[76] His mother filed a lawsuit claiming the police "acted recklessly and deliberately" by killing Albers while he was "simply backing his mom's minivan out of the family garage."[77]

On May 27, 2018, former Army intelligence analyst Chelsea Manning posted two concerning tweets suggesting she might attempt suicide.[78] A wellness check was initiated when people saw the tweets and contacted police.[79] Surveillance cameras in Manning's apartment building recorded the event; the video shows three officers enter Manning's apartment with guns drawn while one officer enters pointing a Taser.[80] The video illustrates how a suicide-related wellness check can escalate to a show of force by police without provocation by a suicidal individual.[81]

---

70. *Id.*

71. *Id.*

72. Ashley Southall, *Naked Black Man Fatally Shot by White Police Officer in Georgia,* N.Y. TIMES (Mar. 9, 2015), https://www.nytimes.com/2015/03/10/us/naked-black-man-fatally-shot-by-white-police-officer-in-georgia.html.

73. Joe Robertson and Tony Rizzo, *FaceTime suicide threat led police to OP student's home before officer shot him,* KAN. CITY STAR (Jan. 22, 2018), https://www.kansascity.com/news/local/article196001754.html.

74. *Id.*

75. *Id.*

76. Joe Robertson et al., *Lawsuit: Teen killed by Overland Park police was 'simply backing his mom's minivan,'* KAN. CITY STAR (Apr. 17, 2018), https://www.kansascity.com/news/local/crime/article209113834.html.

77. *Id.*

78. Micah Lee and Alice Speri, Police Broke Into Chelsea Manning's Home with Guns Drawn—In a "Wellness Check," INTERCEPT (Jun. 5, 2018), https://theintercept.com/2018/06/05/chelsea-manning-video-twitter-police-mental-health/.

79. *Id.*

80. *Id.*

81. *Id.*

The wellness checks described above were performed in the United States. In other regions, such as the Middle East and Southeast Asia, police response may be more unpredictable, and wellness checks may result in criminal penalties such as fines and incarceration. Facebook has deployed its suicide prediction system in nearly every region in which it operates except in the European Union. In some countries, attempted suicide is a criminal offense. For instance, in Singapore, where Facebook maintains its Asia-Pacific headquarters, suicide attempts are punishable by imprisonment for up to one year. Attempted suicide is also illegal in nearby Malaysia, Myanmar, and Brunei.[82] In Islamic countries such as Saudi Arabia, Shari'ah law forbids suicide, which is considered a criminal act.[83] In these countries, Facebook-initiated wellness checks might result in criminal prosecution and incarceration.

The above examples illustrate how social suicide prediction is analogous to predictive policing.

If Facebook's AI misclassifies a user as suicidal, police could be sent to the person's home, which could escalate the situation and provoke a violent confrontation, involuntary hospitalization, or incarceration. Once police arrive following a report that a person is at high risk for suicide, it may be difficult to convince them to leave without being detained. In one case in Ohio, police detained a woman after Facebook warned law enforcement that she might be suicidal.[84] When police arrived, the woman denied having suicidal thoughts, but the officers informed her she would be transported to a hospital against her will if she refused to comply.[85]

B. Privacy Risks

The privacy risks of suicide prediction stem from how prediction data is stored and where the information flows after predictions are made. The risks include leaking of sensitive information through data breaches, and the transfer or sale of personal data to third parties such as data brokers, lenders, employers, and insurance companies. Sale of suicide-related data to these groups can result in stigmatization, exploitation, and discrimination against people categorized as high risk regardless of whether those categorizations are accurate. For instance, a

82. Brian L. Mishara and David N. Weisstub, *The Legal Status of Suicide: A Global Review,* 44 INT'L J. L. PSYCHIATRY 54, 55 (2016).

83. Mohammed Madadin et al., *Suicide Deaths in Dammam, Kingdom of Saudi Arabia: Retrospective Study,* 3 EGYPTIAN J. FORENSIC SCI. 39, 40 (2013).

84. Natasha Singer, *Risks in Using Social Media to Spot Signs of Mental Distress*, N.Y. TIMES (Dec. 26, 2014), https://www.nytimes.com/2014/12/27/technology/risks-in-using-social-posts-to-spot-signs-of-distress.html.

85. *Id.*

life insurance company might purchase suicide prediction data on consumers, and then deny them policies or charge them higher rates than individuals with lower suicide risk scores. In 2017, the U.S. Department of Housing and Urban Development (HUD) filed a complaint against Facebook alleging the company violated the Fair Housing Act by allowing advertisers to exclude people with disabilities, and members of some religious faiths and minority groups, from receiving housing-related ads.[86] Suicide risk scores could similarly be used to deny people access to housing, employment, and other resources, which might further marginalize this already vulnerable population.

In the healthcare system, HIPAA protects patient privacy, and suicide-related data cannot leave the system without first being de-identified. Healthcare providers are also prohibited from sharing non-anonymized health information with third-party advertisers. Thus, medical suicide predictors cannot legally share individualized suicide predictions for marketing purposes. However, because most social suicide predictors are not covered entities under HIPAA, their suicide predictions can be shared with third-parties without first being de-identified, and there are no restrictions on how those predictions may be used. To its credit, Facebook claims its suicide predictions are never used for advertising. However, as the company becomes embroiled in one privacy scandal after another, it may be increasingly difficult for consumers to take the company at its word. Regardless, Facebook is one of many companies making mental health and suicide predictions. Without industry-wide scrutiny and stronger regulation, there will be ample opportunities for abuse.

C. Autonomy Risks

As described above, last year Facebook allegedly enabled advertisers to discriminate against minorities and people with disabilities by excluding them from receiving housing ads. As tech companies increasingly shape people's experiences online and in the real-world, they make decisions on their behalf, potentially depriving them of some degree of autonomy.

One side effect of suicide predictions is that people labeled high risk for suicide may be denied personal and professional opportunities, and in some cases, they may be deprived of civil liberties. The following sections describe how people labeled high risk for suicide may be deprived of opportunities to express themselves on internet platforms and how their Fourth Amendment rights

86. Mason Marks, *Suicide Prediction Technology is Revolutionary. It Badly Needs Oversight,* WASH. POST (Dec. 20, 2018), https://www.washingtonpost.com/outlook/suicide-prediction-technology-is-revolutionary-it-badly-needs-oversight/2018/12/20/214d2532-fd6b-11e8-ad40-cdfd0e0dd65a_story.html?utm_term=.2f4c99f2a344.

may be violated through warrantless searches based on opaque suicide predictions.

1. Censorship

Increasingly, platforms like YouTube, Twitter, and Facebook serve as 21st Century equivalents of the town square where people traditionally gathered to share ideas.[87] Internet platforms go to great lengths to moderate online conversations and maintain civility.[88] They have detailed community guidelines that govern what people can and cannot say, and users are routinely censored or banned for violating the rules.[89]

The New York Times recently described how Facebook's global speech rules are made: "Every other Tuesday morning, several dozen Facebook employees gather over breakfast to come up with the rules, hashing out what the site's two billion users should be allowed to say." Facebook distributes its speech guidelines to about 15,000 content moderators that it employs globally.[90]

According to reports from some moderators, they have mere seconds in which to decide whether content is permissible or objectionable, which makes offloading some of the burden onto AI a necessity.

With over two billion users worldwide, Facebook's guidelines allow it to exercise significant control over global speech. According to its community standards, moderators remove "content that encourages suicide or self-injury, including real-time depictions that might lead others to engage in similar behavior." However, these guidelines are applied inconsistently, and users have little recourse if Facebook removes their content.[91] Some users report having suicide notes removed from the platform without their permission while others report difficulty having them removed.[92]

In 2017, fourteen-year-old British teen Molly Russell killed herself.[93] In

87. Zeynep Tufekci, *Twitter Has Officially Replaced the Town Square,* WIRED (Dec. 12, 2017), https://www.wired.com/story/twitter-has-officially-replaced-the-town-square/.

88. Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech,* 131 Harv. L. Rev. 1598 (2018).

89. Max Fisher, *Inside Facebook's Secret Rulebook for Global Political Speech,* N.Y. TIMES (Dec. 27, 2018), https://www.nytimes.com/2018/12/27/world/facebook-moderators.html.

90. *Id.*

91. *See* Ariana Tobin et al., *Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up,* PROPUBLICA (Dec. 28, 2017), https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

92. *See, e.g.*, Deleting a suicide note? https://www.facebook.com/help/community/question/?id=1572559226321952.

93. Mile Wright and James Cook, *Sir Nick Clegg Says Facebook has Saved 'Thousands' from Suicide,* TELEGRAPH (Jan. 28, 2019), https://www.telegraph.co.uk/news/2019/01/28/sir-nick-

2019, her father publicly claimed Instagram helped kill his daughter by failing to censor content that promotes and glorifies suicide.[94] In response to the story, British Secretary of State for Health Matt Hancock suggested Parliament could ban Internet platforms that fail to remove harmful content from their sites. Meanwhile, facing mounting pressure to improve the fairness of its content moderation, Facebook announced it would create an external board of independent experts to review its "most challenging content decisions."[95] Facebook promises the board will be composed of experts with experience in safety, privacy, and civil rights.[96]

The public health effects of censoring self-harm and suicide-related speech are unknown.[97]There is some evidence suggesting that increased media coverage of suicides promotes copycats and increases suicide rates.[98] However, it is unclear what effect censoring suicide-related speech on social media has on suicide rates.[99] Unlike the speech of news media, which is protected from government censorship by the First Amendment, the speech of social media users is not protected because Internet platforms and their content moderators are not government entities. Nevertheless, there may be public health arguments for ensuring freedom of expression for users of online platforms.

Though it is possible that uncensored suicide-related speech could inspire copycats, it is equally plausible that stifling public discussion of suicide contributes to its taboo nature and inhibits people from seeking and receiving needed help and support. Somewhat surprisingly, Facebook does not censor suicide-related expression when users live-stream their suicide attempts. Its rationale is that leaving the stream running "until the point of no return" maximizes the chance that viewers of the stream can send for help. The problem is Facebook makes these decisions unilaterally, censoring some instances of suicide-related speech, but not others, and its decisions are not transparent or evidence-based.

---

clegg-says-facebook-has-saved-thousands-suicide/.

94. *Id.*

95. Draft Charter: An Oversight Board for Content Decisions, Facebook, https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-1.pdf.

96. *Id.*

97. Marks, *supra* note 23.

98. Madelyn S. Gould, *Suicide and the Media,* 932 CLINICAL SCI. SUICIDE PREVENTION 200 (2001).

99. Thomas Ruder et al., *Suicide Announcement on Facebook,* 32 CRISIS: J. CRISIS INTERVENTION SUICIDE PREVENTION 280 (2011).

2. Warrantless Searches

As AI-based suicide prediction tools proliferate, they will play an increasing role in police and doctors' decisions to involuntarily hospitalize people for treatment or medical observation. Civil commitment is an intervention that strips people of liberty and autonomy, and it is not without risks.[100] Nevertheless, it is permitted by state laws when individuals are deemed a risk to themselves or others.[101] If a person is deemed high-risk by social suicide prediction tools, prompting police officers to respond to that person's home, and the person does not answer the door, then police could enter the home without first obtaining a search warrant.

In the U.S., the Fourth Amendment protects people and their homes from warrantless searches.[102] However, under exigent circumstances doctrine, police may enter homes without warrants if they reasonably believe entry is necessary to prevent physical harm. Stopping an imminent suicide attempt clearly falls within this exception. However, it may be unreasonable to rely on opaque AI-generated suicide predictions to circumvent Fourth Amendment protections when no information regarding their accuracy is publicly available. As described above, Facebook makes suicide predictions based on internal data rather than data from real suicides. We don't know how accurate its predictions are, what criteria it uses to decide when law enforcement should be contacted, or what information it provides to police. Exceptions to the warrant requirement should not be made based on such paltry information.

III. CONCLUSION

Medical and social suicide prediction tools may be beneficial to individuals and promote public health. However, they also pose a variety of risks to people's safety, privacy, and autonomy. To minimize those risks, new norms and regulations must be developed to control how suicide predictions are made and used. For example, to protect consumer autonomy, suicide prediction methods could be made more transparent, and users could be given unambiguous opportunities to opt-out and delete prediction information; to protect consumer privacy and minimize the risk of exploitation, suicide predictions should not be

100. *See* Megan Testa & Sara G. West, *Civil Commitment in the United States,* 7 Psychiatry 30 (2010).

101. *Id.*

102. Ken Wallentine, *Should I Stay or Should I Go—If You Respond to a Call Involving a Suicidal Person Who's Not Endangering Anyone Else, It Might be Best to Not Intervene*, POLICE MAGAZINE (Oct. 16, 2017), http://www.policemag.com/channel/patrol/articles/2017/10/should-i-stay-or-should-i-go.aspx.

used for advertising or be shared with third parties; and to protect consumer safety and autonomy, "soft-touch" suicide interventions such as referrals to counseling centers, could be prioritized over "firm-hand" interventions such police-mediated wellness checks.

In some cases, healthcare norms and regulations could be imported for use in social suicide prediction. For instance, social suicide prediction research should be approved by independent IRBs, and ongoing suicide prediction programs should be monitored for safety and efficacy by independent data monitoring committees. Though HIPAA does not currently apply to social suicide prediction, to protect consumer privacy, tech companies could voluntarily adopt HIPAA-like standards, or stricter standards could be imposed on them through new privacy legislation.