Philosophy & Technology (2022) 35:103
https://doi.org/10.1007/s13347-022-00599-z

RESEARCH ARTICLE

AI, Suicide Prevention and the Limits of Beneficence

Aurélie Halsband[1] · **Bert Heinrichs**[2,3]

Received: 12 August 2022 / Accepted: 5 November 2022 / Published online: 28 November 2022

Abstract

In this paper, we address the question of whether AI should be used for suicide prevention on social media data. We focus on algorithms that can identify persons with suicidal ideation based on their postings on social media platforms and investigate whether private companies like Facebook are justified in using these. To find out if that is the case, we start with providing two examples for AI-based means of suicide prevention in social media. Subsequently, we frame suicide prevention as an issue of beneficence, develop two fictional cases to explore the scope of the principle of beneficence and apply the lessons learned to Facebook's employment of AI for suicide prevention. We show that Facebook is neither acting under an obligation of beneficence nor acting meritoriously. This insight leads us to the general question of who is entitled to help. We conclude that private companies like Facebook can play an important role in suicide prevention, if they comply with specific rules which we derive from beneficence and autonomy as core principles of biomedical ethics. At the same time, public bodies have an obligation to create appropriate framework conditions for AI-based tools of suicide prevention. As an outlook we depict how cooperation between public and private institutions can make an important contribution to combating suicide and, in this way, put the principle of beneficence into practice.

Keywords Artificial intelligence · Autonomy · Beneficence · Privacy · Social media · Suicide

✉ Aurélie Halsband
ahalsban@uni-bonn.de

[1] German Reference Centre for Ethics in the Life Sciences (DRZE), Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

[2] Institute of Neurosciences and Medicine: Brain and Behavior, Forschungszentrum Jülich, (INM-7), Jülich, Germany

[3] Institute of Science and Ethics (IWE), Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

1 Suicide as a Topic of Philosophy

Suicide is a classic theme of philosophy and questions surrounding its assessment have occupied western philosophy since antiquity. In Plato's dialogue Phaedo, Socrates remarks that no man has a right to take his own life (Plato, 1951, Phaedo 62a-c). In a similar vein, Aristotle calls suicide a cowardly act (1925, EN 1115a-1116a). Other authors of antiquity regard it as permissible under certain conditions (Cicero, 1931, I, 18). In the succeeding Christian tradition, suicide is considered a serious sin (Augustine, 1972, I, 22; Aquinas, 1947, STh II-II q. 64. a. 65). And in the eighteenth century, a liberal defense of suicide was put forward by Hume (2005). Even in the recent philosophical past, suicide has still been thought about intensively. In fact, Camus famously claimed: "There is but one truly serious philosophical problem and that is suicide" (2005, p. 1). Equally famous is Wittgenstein's reflection: "If suicide is allowed, then everything is allowed. If anything is not allowed, then suicide is not allowed" (1984, p. 91).

Apart from this philosophical debate, there is a consensus that most cases of suicide are due to pathological conditions and that it is important to help people overcome this kind of crisis situations (Lutz et al., 2017). Our philosophical analysis is targeted at an evaluation of means which may be used to prevent suicides. Our contribution is neither a reflection on whether suicide is an "elementary sin" (Wittgenstein, 1984) or an act of utmost self-determination. Nor does it formulate a position on whether suicide is a central problem of philosophy (Camus, 2005) or just a side issue. Instead, the scope of our paper is less broad and more practical: Assuming that it would be desirable to reliably discern and help those people who are at risk of suicide or suicidal ideation due to pathological circumstances, should social media data be analyzed by artificial intelligence (AI) for this purpose, and if so, under what conditions?

We focus on algorithms that can identify persons with suicidal ideation based on their postings on social media platforms.[1] Facebook developed such an algorithm and uses it since 2017. In a recent study, the development of a similar algorithm based on Twitter messages was reported (Roy et al., 2020). However, assuming that it would be morally desirable to help people with suicidal thoughts does not automatically imply that private companies like Facebook are justified in using such algorithms as a suicide prevention tool.[2] To find out if that is the case, we start with providing two examples for the use of AI as a means of suicide prevention in social media (Section 2). Subsequently, we develop two fictional cases to explore the scope of the principle of beneficence (Section 3 and Section 4) and apply the lessons learned to Facebook's employment of AI for suicide prevention (Section 5). We then consider whether Facebook's commitment can be considered a meritorious

[1] Our focus is thus on "social suicide prediction tools" which differ from "medical suicide prediction tools". For this distinction cf. D'Hotman & Loh, 2020.

[2] Since 2021 the social network Facebook is owned by Meta Platforms, Inc. For reasons of simplicity and because our analysis focuses on this specific service provided by Meta Platforms, we further refer to Facebook and consider it a private company.

act (Section 6). This leads us to the general question of who is entitled to help (Section 7). Finally, we discuss balancing issues (Section 8) and summarize our findings (Section 9).

2 Using Algorithms for Suicide Prevention on Social Media Data

The advantage of using AI for the identification of persons who are at risk of developing suicidal ideation or who are already expressing such thoughts is AI's ability to investigate large sets of data (Bernert et al., 2020). Regarding suicide prevention, this provides the opportunity to include a wide array of factors which may raise the risk of suicidal ideation for specific persons. Also, it opens further opportunities to discover new patterns of factors which stand in relation to suicidal ideation. In sum, the hope is to improve the accuracy of prediction in the context of suicidal ideation by the introduction of AI technologies which "[…] hold the potential to impact suicide on broad scale" (Bernert et al., 2020, p. 1). Social media is an area particularly suitable for the identification of potentially suicidal individuals by means of AI: Data is easily accessible for evaluation by researchers or platform providers and it often contains intimate details of personal nature. Several studies have already shown the potential of such AI applications (D'Hotman & Loh, 2020). Social networking services such as Twitter or Facebook are frequently used by younger persons where suicide is one of the major causes of death. In 2019, suicide was globally the fourth leading cause of death in young people aged 15–29 years for both sexes (World Health Organisation, 2021). What is more, younger persons have been shown to frequently choose online media to express suicidal ideation (cf. i. a. Pourmand et al., 2019). Hence, using AI may provide new opportunities of suicide prevention especially targeted at younger persons (Roy et al., 2020; Sueki, 2015).[3]

Identifying potentially suicidal persons based on their social media posts is considered an up to now rather novel approach representing a "unique promise" to impact suicide prevention (Bernert et al., 2020, p. 18). Various types and uses of AI in the social media context share the goal to detect persons that are either at risk of developing suicidal thoughts or are already expressing them. They differ, among other things, in whether they also aim at providing help (Gomes de Andrade et al., 2018).

To illustrate the different AI approaches to suicide prevention in the context of social media (e. g., Coppersmith et al., 2018; O'Dea et al., 2017), we briefly present two examples. The *Suicide Artificial Intelligence Prediction Heuristic (SAIPH)* has been developed by an independent research team and can be applied to

[3] The use of AI on social media data could, of course, be attractive for other purposes as well, especially for the fight against crime and terrorism. In fact, AI is already being used for these purposes, although it is controversial, among other things, from an ethical point of view. However, due to its monopoly on the legitimate use of force, the fight against crime and terrorism is the exclusive responsibility of the state. One of the main ethical questions regarding private companies such as Facebook, therefore, is how far they can be compelled to cooperate with security authorities to help combating unlawful activities. In contrast, in the context of a private company's voluntary commitment to prevent suicides by using specific algorithms, one of the main questions is up to which point states should welcome and embrace this.

publicly available Twitter posts. Our second example is the set of suicide prevention algorithms that have been developed by Facebook and are currently used by the company to screen users within its specific social network.

SAIPH uses machine learning techniques to reveal patterns within publicly available posts (tweets) of persons on Twitter which indicate a risk of developing suicidal ideation. Moreover, the developers of SAIPH also claim to provide information about persons at risk before they express those intentions (Roy et al., 2020). As a means of prevention, the developers hold that SAIPH can help detecting who and when someone will be at an especially high risk of developing suicidal thoughts within the following 10 days (Roy et al., 2020). SAIPH was developed using publicly available tweets of persons collected over 2 years. In the training stage, a series of neural networks was generated which evaluated the text-based posts for key words (proxies) indicating psychological states such as anxiety or depression which are known to be linked to risks of developing suicidal thoughts. To provide the algorithm with training data on the occurrence of such psychological states and how they relate in time to the potential later development of suicidal ideation, tweets of persons that expressed suicidal ideation in the past were integrated as well as those of a control group without such ideation. This information on suicidal ideation and its causal as well as temporal relation to psychological states was then fed into a so-called random forest model which allows for classifying new cases of persons at risk of suicidal ideation based on their tweets. SAIPH's ability to predict the risk of suicidal ideation has been validated on regional suicide rates and on cases of celebrities which are known to have committed suicide, and which were active on Twitter before their death (Roy et al., 2020).

Facebook is also using a random forest model for suicide prevention (Gomes de Andrade et al., 2018). However, the training data differs: Facebook used reports of potential cases of suicidal ideation. These reports are part of Facebook's strategy to address suicides before their occurrence, providing members the opportunity to contact a dedicated review team whenever they want to raise awareness to posts of others that may indicate suicidal thoughts (Gomes de Andrade et al., 2018). The posts that have been evaluated by the review team served as a pool of affirmed cases of posts indicating suicidal intentions as well as of false positive cases.[4] Subsequently, this set of data was used to train an algorithm for classifying posts and comments based on key words associated with suicidal ideation by n-gram linear regression. To improve text evaluation quality, DeepText was later included as a further element. In addition to text, Facebook successively included other types of information, e.g., on the time of day of posts, on the type of content (e.g., video or plain text) and on friends' reactions to these posts as supplements for risk measurement (Gomes de Andrade et al., 2018). The model created this way is now used to extract different

[4] The precise criteria for the evaluation of such reports as either affirmed cases of suicidal ideation or false positive reports by Facebook's review team have not been published (cf. i. a. Singer 2018). Facebook rather repeatedly refers to its exchange with external experts on suicide prevention in regard to the development and evaluation of its suicide prevention procedures including AI (cf. e.g., Gomes de Andrade et al., 2018, p. 676).

features of posts, focusing on their use as a proxy for suicidal thoughts. In addition to the algorithm's support in *identifying* posts potentially indicating suicidal ideation, it is also being used to optimize the evaluation of posts reported by Facebook members. In a process called "queue prioritization" (Gomes de Andrade et al., 2018, p. 674), the algorithm pre-selects those previously reported posts that show features of a specifically acute danger of suicidal ideation. The algorithm accordingly assigns higher priority to these and is thus targeted at making the evaluation by the review team more efficient in terms of suicide prevention. Facebook combines a proactive assessment of posts potentially indicating suicidal ideation with a response mechanism that includes the communication of helplines to users which have issued such posts (Gomes de Andrade et al., 2018). In cases where an attempt to suicide seems to be imminent, Facebook's proactive response includes contacting local emergency services.[5] Twitter itself has a similar system for reporting and evaluating posts. It does, however, not use AI to identify such posts and is therefore only responsive to reports of its members. Although SAIPH can be used on Twitter data, Twitter itself has neither initiated the development of such an algorithm nor has the company up to now publicly announced that it plans to use such AI-based suicide prevention tools on its platform in the future. While the responsive approach is less problematic, it is the proactive use of such algorithms to identify potentially suicidal persons that raises a whole series of difficult ethical questions, e.g., who should be allowed to use such algorithms, under what conditions, and for what purposes. A good starting point for an in-depth ethical analysis is the concept of beneficence.[6] After all, suicide prevention is arguably about helping others. The above questions, then, are questions about the scope of the principle of beneficence.

3 An Initial Case: Help with Acute Suicide Risk and the Principle of Beneficence

Let us imagine the following case: Alice crosses a deserted bridge. She thinks she is alone and is contemplating her thoughts when she suddenly spots another person — Bob — standing behind the bridge railing. Alice is immediately aware that Bob is considering jumping off the bridge. She feels an instant obligation to help. Alice is initially unsure about what exactly to do, but after a short moment of contemplation, she approaches the person and exclaims: 'Hey, what are you doing? Wait, don't jump!' Alice enters a conversation with Bob, in the course of which Alice is able to convince him to come back to the safe side of the bridge railing. Alice and Bob talk for a long time and eventually Alice takes him to a psychiatric facility where he

[5] A controversial case of this type of emergency service notification by Facebook was reported in 2018 by the New York Times. As a result of Facebook's notification, a woman that had been identified by Facebook's suicide prevention procedure as being at risk of suicide was forced by a local police officer to undertake a mental health check-up although she repeatedly asserted not to intend self-harm (Singer, 2018).

[6] We limit our analysis to ethical arguments and leave legal considerations aside. The latter can be found, for example, in Celedonia et al. (2021).

is hospitalized. In this case, most people would agree that Alice ought to help and that she cannot go on and pretend that none of this concerns her. To be sure, it is not clear how far this obligation extends and what exactly it entails. For example, Alice certainly does not have to put herself in great danger to save Bob. It seems, however, clear that Alice is overall morally obligated to help. This is because Alice's moral obligation is rooted in the principle of beneficence which is one of the most basic ethical principles. Along with the principle of autonomy, the principle of non-maleficence, and the principle of justice it forms a group of broadly accepted ethical principles.[7] We will not attempt to justify the principle of beneficence here, but simply assume that it is a basic ethical principle which can guide the ethical evaluation of the different uses of AI (cf. for a similar methodological approach to AI ethics Floridi et al., 2018). In their influential work on bioethics, Tom Beauchamp and James Childress provide a characterization of the principle of beneficence. According to them, it includes, among other things, the following *prima facie* rules:

1. Protect and defend the rights of others.
2. Prevent harm from occurring to others.
3. Remove conditions that will cause harm to others.
4. Help persons with disabilities.
5. Rescue persons in danger. (Beauchamp & Childress, 2019, p. 219)

This characterization elucidates how the principle of beneficence is relevant to the situation described earlier. It explains why Alice felt immediately obligated or, if that had not been the case, why we would have morally reproached her. Intervening in acute suicide attempts is certainly a case of acting to prevent harm from occurring to someone else (Rule 2), as well as with saving a person in danger (Rule 5).

As already indicated, however, the exact scope and conditions for the application of the principle are complicated and often controversial. For instance, there would be no obligation to help for Alice if Bob were to seriously injure her. Also, the situation would be more difficult if Bob repeatedly and credibly asserted that he did not want Alice's help. Beauchamp and Childress address issues around the applicability and conditions of the principle of beneficence by providing a list of conditions that must be satisfied to assign obligations of beneficence:

[…] [A] person X has a prima facie obligation of beneficence, in the form of a duty of rescue, toward a person Y if and only if each of the following conditions is satisfied (assuming that X is aware of the relevant facts):

[7] The best-known and most elaborate version of an ethical framework operating with these four principles was developed by Tom Beauchamp and James Childress (2019). They originally conceived it for the context of biomedicine. However, the framework can be — and indeed often is — applied in other contexts as well. Obviously, autonomy, non-maleficence, beneficence, and justice are very general ethical principles that are not unique to biomedicine. In what follows, we adopt from Beauchamp and Childress only this general framework and some further considerations that do not depend on the biomedical context.

1. Y is at risk of significant loss of or damage to life, health, or some other basic interest.
2. X's action is necessary (singly or in concert with others) to prevent this loss or damage.
3. X's action (singly or in concert with others) will probably prevent this loss or damage.
4. X's action would not present significant risks, costs, or burdens to X.
5. The benefit that Y can be expected to gain outweighs any harms, costs, or burdens that X is likely to incur. (Beauchamp & Childress, 2019, p. 222)

In the paradigmatic case of Alice and Bob, all five conditions are fulfilled: Bob's attempt to commit suicide significantly endangers his life. Alice, on the contrary, faces only the minor burden of talking to Bob which clearly weighs less than the benefit to Bob. Also, Alice can safely assume that her intervention is required since she is the only person around to help Bob. So, let us assume that there is an obligation for Alice to help Bob and that the conditions for doing so are reasonably clear.

Does this suggest anything for the use of AI for suicide prevention? Is there a comparable obligation to provide help in this case as well? If so, who does it affect? To be sure, when it comes to the use of AI, many aspects are different. First, there is no immediate encounter like there is between Alice and Bob. Second, Alice did not go to the bridge to help Bob but passed by purely by chance. Third, Alice can help without further preparation, whereas suicide prevention by means of AI requires prior financial and technological investments to develop and to run the algorithms before the occurrence of a suicidal act.

It may help to consider a second case that lies, in a sense, between the case of Alice and Bob, in which the ethical circumstances are clear, and the case of AI and suicide prevention.

4 A Second Case: Obligations to Help and Limits of Reasonableness

Let us assume that Dave has a company that is in charge of bridge maintenance. In order to detect safety-relevant changes on bridges at an early stage, Dave's company has installed camera systems on the bridges. These camera systems send real-time images to the company's headquarters, where they are monitored. In the event of any anomalies, the company sends engineers to the bridges to carry out on-site inspections. It happens that people can be seen on the camera images who are crossing the bridge railing. Of course, the camera images are not very conclusive, but it could be that these are people who want to jump off the bridge to commit suicide. Imagine Dave sitting in front of the screen in his office and seeing Carol climb over the railing and stop there, looking down into the depths, apparently about to jump. Does Dave have an obligation to intervene?

The framework of Beauchamp and Childress is once again useful for an initial analysis. As far as the expected action is concerned, the cases are similar: Since Carol is standing behind the railing, it can be assumed with high likelihood that she is about to commit suicide, thus endangering her life. However, the burdens for Dave

and his bridge maintenance company differ. As with Alice, Dave enters the situation in which a person needs help only by coincidence. His company's cameras show persons behind the bridge railing as a side-effect. Therefore, to be able to *actively* help, Dave would need to make sure that staff is available and can be sent to the bridges in cases of emergency. In contrast to Alice, the company's ability to help depends on a dedicated investment in such a rescue system. Most would probably agree that this is not something one can expect from a private company. What can be expected, however, is that Dave calls the rescue services in an emergency. In other words, he has the same obligations as any other citizen. His company's camera system merely results in him potentially observing suicide attempts more frequently and, therefore, probably getting into situations where he has to call rescue services more often. Of course, there may be borderline cases. It could be, for example, that Dave's company's cameras would have to be set up at a special angle to capture suicidal people well. If this would not interfere with bridge surveillance, then such a measure could be considered reasonable. But if Dave had to install more or completely different camera systems, then it would be more appropriate to say that it was the state's responsibility to provide such systems, while Dave's company might have to allow these public cameras to be placed on the company's carriers. Such casuistic considerations can be important in practice and finding convincing trade-offs can be difficult. However, these considerations do not change what has already been observed in view of the general scope of the obligations: Dave and his company have no obligation to permanently monitor the bridges for suicides, and they have no obligation to actively provide help in an emergency. If they happen to observe suicides, they are reasonably only obligated to contribute to providing help by notifying the emergency services.

5 A Real Case: Facebook's Employment of AI for Suicide Prevention

How does this compare to the use of AI for suicide prevention by a private company like Facebook? Like Dave and his company with its camera devices, Facebook is in a genuine position to identify persons at risk of suicide by means of AI. Facebook, too, might therefore have some obligation to contribute to suicide prevention. What is more, regarding the cost–benefit-ratio in Beauchamp's and Childress's framework, Dave's company and Facebook are also in a comparable situation. While the effort to organize help in acute suicide cases, for example by alerting the emergency services or informing public crisis intervention centers, seems reasonable, it would be asking too much of both Dave's company and Facebook if they were to provide this help themselves.

Unlike Dave and his company, however, Facebook seems willing to take on the costs associated with more extensive help. Facebook's CEO Mark Zuckerberg frames the effort of the company as an attempt to build a global and safe community, providing an "infrastructure to give their friends and community [i.e., the Facebook community and its single members] tools that could save their life" (Gomes de Andrade et al., 2018, p. 678 note 644). He presents Facebook's use of AI for suicide prevention as an act between "Facebook friends" who stand in a special

moral relationship to each other, which entails more extensive moral obligations. Note that Beauchamp and Childress start their framework with the following parenthesis: "Apart from close moral relationships, such as contracts or the ties of family or friendship, we propose that a person X has a prima facie obligation of beneficence if …" (Beauchamp & Childress, 2019, p. 222), enumerating the above quoted conditions. However, the moral concept of friendship is different from the one implied by Facebook, inter alia, because it describes shared activities and significant direct interactions between the persons involved to qualify as friendship (cf. Helm, 2021). Yet, these conditions can hardly be fulfilled for every member of the Facebook community which comprises millions of persons. Hence, the relation between Facebook friends hardly qualifies as a special moral relationship entailing more extensive obligations to help.

Another difference between Dave and his company's contribution to suicide prevention and Facebook's engagement for this purpose may be brought forward to justify a differing categorization: In contrast to bridge surveillance the use of social media seems to have a causal relation with the development of suicidal ideation especially regarding adolescents. If the use of Facebook as a social media platform promotes suicidal ideation within the group of younger persons in some cases, would this justify a special obligation for social media platform providers to offer help? If such a connection existed, then this could indeed be the case. However, empirical evidence does not support this line of argument. A link between an increased suicidal ideation and social media use is established only in the context of "problematic" social media and internet use (Sedgwick et al., 2019, p. 540; for a slightly different assessment see Celedonia et al., 2021). Moreover, in cases where social media use is correlated with suicidal ideation, scientific evidence is lacking as to whether social media use in general or other correlated circumstances "such as sleep disturbance and cyberbullying" may represent "confounders" (Celedonia et al., 2021, p. 3). Finally, some types of social media use may even decrease suicidal ideation because persons in crisis digitally reconnect with others. To be clear, this does not suffice to refute a potential link between social media use and suicidal ideation. Rather, it stresses that the sparse scientific evidence currently available does not suffice to justify a special moral obligation of beneficence for social media platform providers in the context of suicide prevention.

However, focusing on cases of imminent suicide attempts may have distracted from a genuine feature of Facebook's involvement in suicide prevention, namely its access to data which can be used to identify persons at risk of suicide *at a very early stage* of ideation. At first sight, the potentially imminent action of jumping off the bridge in the cases of Alice and Bob as well as in the case of Dave and Carol may instantiate more directly a suicidal intention than a conglomerate of posts published online. Among other things because standing behind the railing of a bridge is already an integral part of the act of taking one's own life whereas social media posts remain in the preliminary stages of such an act. At closer inspection, this imminence of an act does not exclude situations with less imminent character from the principle of beneficence. This is because the overall goal in suicide prevention is not only to impede attempts but to assist persons in crisis situations in such a way that they neither pursue nor develop the wish to end their lives. Accordingly, intervening *before*

imminent attempts is even more pressing with regard to the objective of reducing the risk of suicide ideation overall. And this capacity for early intervention is Facebook's special asset in using AI for suicide prevention. Therefore, if Facebook is outstandingly suited to provide very early access to persons at risk of suicide by the employment of AI, does this imply an obligation of beneficence?

As with Dave and his company, Facebook does not satisfy all five conditions for the assignment of an obligation to help. Although Facebook is willing to accept the costs required for being able to actively help, this does not suffice to satisfy the conditions regarding the cost–benefit-ratio of an *obligation* to help. Most importantly because it cannot be expected in general from a private company to face this kind and scope of costs. Moreover, as noted above, the loose concept of friendship between Facebook members does not justify the assignment of a special moral relationship supporting specific obligations of beneficence. Finally, neither possible correlations between social media use and the development of suicidal ideation nor Facebook's exclusive access to persons at very early stages of suicidal ideation alone justify the assignment of obligations of beneficence. In conclusion, Facebook has no obligations to proactively help people with suicidal ideation. All that can be expected is that they inform the relevant authorities if they become aware of cases.

6 Facebook's Commitment as a Meritorious Act

Currently, Facebook is doing more. The company does not only use an existing infrastructure and sound the alarm on randomly observed cases. It spends considerable effort on developing suitable algorithms, routinely uses them to search through huge amounts of data, and points out offers of help in the case of a member's activity potentially indicating suicidal ideation. How should this commitment be evaluated against the background of the previous considerations?

Facebook's approach could be seen as a meritorious (sometimes also called supererogatory) act.[8] This is generally understood to mean actions that are not strictly morally required but go far beyond. We know such acts of merit from everyday life, and it is mostly about help that is given without being strictly demanded. Should not we just applaud Facebook for going to such lengths to combat a serious societal problem — suicides?

[8] Of course, it is important to keep in mind that Facebook's involvement in this area is likely to benefit the company's reputation. Thus, strategic market considerations could also have played a role in the development and implementation of the algorithms. In addition, the company could also try to gain a foothold in the extremely lucrative medical market in this way. One could therefore regard our approach as naive and rule out from the outset that Facebook's involvement is a meritorious act. However, we are by no means gullible but rather focus on the help for suicidal people that Facebook provides and ask how this is to be assessed. We limit our discussion to ethical aspects of AI-based suicide prevention provided by private companies and do not aim at providing an overall ethical assessment of Facebook. Regarding Facebook's initiative, we do not rule out the possibility that there are other considerations that should be considered in an overall assessment.

To that end, let us look again at Dave and his company. Let us assume that the original camera system is not particularly good at detecting suicidal people. One day, however, Dave unambiguously spots a person with suicidal intent through the camera system. Let us further assume that he takes this as an opportunity to install a much more elaborate system at his own expense and to assign people in his company to guard the monitors around the clock. What is more, in an emergency Dave does not alert the emergency service, but regularly sets out on his own. Would this be considered a meritorious act? Or would it be more accurate to say that Dave is overdoing it? More than that, would this perhaps indicate that meritorious action is in danger of becoming a problematic action here?

In retrospect, most persons do not want many people to know about a previous suicide attempt. Public agencies are therefore subject to confidentiality, and even rescue workers such as firefighters are not allowed to simply report on very private issues that they come to know about during their missions. Apart from the fact that rescue workers have the competencies necessary to provide assistance in emergencies, they also belong to the public sector, for which special rules apply. Rules which provide limits to who, when, and how far actions intended as help are legitimate. These rules primarily stem from the discussion about legitimate paternalistic acts, in which the respect for the autonomy of persons is overruled by the principle of beneficence. In light of this, Dave ought to have contacted public officials when he realized that his cameras might help save lives. Together, they could have decided on appropriate measures. Dave could have contributed his knowledge of bridges and cameras, and public agencies could have used this knowledge to improve emergency services for persons at risk of suicide. This way, a private individual like Dave would not have suddenly been drawn into a highly sensitive area and would not be at risk of violating the privacy of others.

This reasoning can also be applied to Facebook. With regard to the issue of exceeding competencies, the case of Facebook is even more pressing than in the example of Dave and his company. Facebook evaluates huge amounts of data and encounters highly sensitive information. And although users have consented to this by accepting the company's terms and conditions, this does not exclude the possibility that specific actions can still be problematic. The use of algorithms for suicide prevention could be such a case. If Facebook's infrastructure is capable of reliably detecting suicide cases, and if it really is possible to prevent such cases in this way, then it is indeed a good thing. However, this does not automatically imply that Facebook itself should operate these algorithms. Rather, it becomes clear, that the questions of who is obligated to help and who may meritoriously do so need to be addressed in combination with another question: Who is *entitled* to provide help?

7 Who Is Entitled to Help?

Who could oppose actions targeted at promoting beneficence? If a person or institution is willing to provide help to persons in a crisis leading to suicidal thoughts or even actions, should not we simply appreciate it? And should not we acknowledge this even more when it is done beyond any moral obligation? It is usually assumed

that a person's or institution's engagement for beneficence reaches a limit when it conflicts with other important moral values such as the already mentioned core ethical principles of non-maleficence, autonomy, and justice. In the context of suicide prevention, such conflicts often arise with the principle of autonomy because measures aimed at helping people in crisis may simultaneously constitute an interference with their self-determination. This conflict within suicide prevention is often framed as the need to weigh the principle of beneficence against the privacy of a person (understood as a specification of the principle of autonomy). In more general terms and referring to a definition of paternalism provided by Beauchamp and Childress, intervening with the aim of suicide prevention may be described as "the intentional overriding of one person's preferences or actions by another person, where the person who overrides justifies the action by appeal to the goal of benefiting or of preventing or mitigating harm to the person whose preferences or actions are overridden" (Beauchamp & Childress, 2019, p. 231f.)

The intentional overriding of actions by one person with the intention of benefiting another person captures a plethora of actions involved in suicide prevention. Remember, however, that we presupposed that there are many suicide attempts characterized and caused by pathological conditions. These suicide attempts do not represent an autonomous choice which ought to be respected. Even more, they severely restrict the autonomy of the person. In other words, we assume that some types of suicide prevention can be considered legitimate paternalistic acts. Accordingly, the central question is what requirements these paternalistic acts must meet to be legitimate.

At first sight, the definition of paternalism and the included conditions for actions aiming at helping others support Facebook's engagement for suicide prevention. After all, Facebook appeals to the safety of its community. Yet, at closer inspection respect for autonomy calls for further constraints. Beauchamp and Childress consider such constraints in the context of health professionals and their interference with patient autonomy, two of which are especially relevant for the present context: Only those actions may qualify as legitimate paternalistic acts which, first, have "no morally better alternative to the limitation of autonomy that will occur" and, second, which are the "least autonomy-restrictive alternative" (Beauchamp & Childress, 2019, p. 239).

Note that the employment of AI for suicide prevention discussed above is primarily targeted at identifying and contacting potentially suicidal persons. AI is not used to actively provide help. These further measures are delegated to emergency services or other professional institutions. With respect to the Facebook case, then, the central question is: Does it constitute an illegitimate restriction of privacy if Facebook contacts a person previously identified as suicidal by an AI that analyzed personal social media data? Is it justified if they call emergency services?

Consider the first constraint: Is there a better alternative to the limitation of autonomy that will occur? Probably not. The use of AI presents unprecedented opportunities and could help save many lives: A central asset of the employment of AI for suicide prevention is that it can be used to identify those persons at risk of suicide ideation who are already being assisted in a clinical context but do not express their ideation in this context and those who are not presently in a clinical context but at

risk of suicidal ideation. In both contexts, research suggests that the employment of AI on social media data significantly improves the identification of persons at risk. Especially with regard to suicide prevention for adolescents, AI-based evaluations of social media activities is a genuinely promising tool: as mentioned above, suicide is one of the major causes of death within this sub-group of young persons who also often do not express their suicidal thoughts in front of medical personnel but very frequently online (Bernert et al., 2020; Pourmand et al., 2019; Roy et al., 2020; Sueki, 2015). Yet, a final assessment of potential alternatives to the limitation of autonomy requires the possibility to scientifically assess the algorithm's performance in correctly identifying persons with suicidal ideation, especially with regard to the criteria which were used to train the algorithms (Celedonia et al., 2021; Marks, 2019). After all, alternatives can only be assessed if their efficiency can be evaluated. To assess whether alternatives are available requires access to information about the AI's efficiency in comparison to other tools of suicide risk assessment such as the well-established questionnaire-based Scale for Suicide Ideation (SSI) or the Beck Scale for Suicide Ideation (BSS), i.e., the self-report version of the interviewer-administered SSI. This includes testing whether the algorithms work equally well for different groups. It is known from past experience that algorithms of this type often exhibit strong biases that lead to significantly worse results, e.g., for minorities. A high false-positive rate could cause disproportionate harm to individual groups (Celedonia et al., 2021). This could mean that the advantages and disadvantages, including privacy intrusions, of such an algorithm could be very unequally distributed between different groups, which could possibly speak against their use under considerations of fairness. Taken together, this suggests that Facebook should disclose information about its algorithms for independent evaluation. This is to say, applying AI for identifying persons with suicidal ideation should be open for external revision before qualifying as a legitimate paternalistic intervention in a person's privacy. This is a prerequisite for a comprehensive assessment of AI-based suicide prevention that considers both benefits and costs including privacy protection. Note that such an assessment does not necessarily include a full disclosure of the algorithms to the public. It may be sufficient to allow access to the competent authorities as has long been the case, for example, in the testing of new pharmaceuticals.

Now consider the second constraint: Is there a less autonomy-restrictive way to identify persons at risk of suicidal ideation? The above considerations suggest that AI-based evaluation of social media activities ought to be conducted by institutions and persons operating under specific rules of confidentiality. In other words, although it may be a legitimate paternalistic act to identify persons at risk by their social media activities with the help of AI not everyone is entitled to do so. Currently, AI evaluations of social media activity are primarily applied in clinical or emergency contexts (cf. i. a. Bernert et al., 2020). In these contexts, strict rules of professional confidentiality apply which regulate the collection, use and storage of data, safeguarding the patient's privacy and regulating the personnel involved (Beauchamp & Childress, 2019, p. 242ff.). In fact, current research additionally argues for an extension of these specific professional rules to encompass autonomy-related challenges by the use of AI (Laacke et al., 2021). In contrast, in countries like e.g., in the USA (cf. Celedonia et al., 2021), these rules of confidentiality do not apply

to private companies outside the health care sector such as Facebook even when they act *as if* they were health care providers (Celedonia et al., 2021, p. 8; Barnett & Torous, 2019). Consequently, suicide prevention with the help of AI by private companies does only qualify as legitimate paternalistic act if it meets the conditions of confidentiality sketched out above.

In a similar vein, Celedonia et al. (2021) call for embedding companies that use AI for suicide prevention in a fiduciary framework. They suggest building this framework for the use, access and storage of sensitive personal health data in analogy to medical research projects, involving an ethical review process and procedures guaranteeing informed consent. With the specific focus on Facebook, a less restrictive way to identify persons at risk of suicidal ideation may require inter alia adding a passage in the company's terms and conditions for users in which they explicitly accept that Facebook may contact local emergency services whenever they have been identified as a person at imminent risk of suicide (cf. for a critical assessment Celedonia et al., 2021; Marks, 2019). Moreover, as with the rules that apply to individuals working in emergency services or health care in general, the Facebook suicide review team should also be bound by confidentiality rules that require its members to, among other things, keep the names of individuals identified by the AI as having suicidal ideation secret from others. Ultimately, a private company's general handling of sensitive data would need to be considered, in particular in view of the non-authorized disclosure of sensible information to third parties.

An alternative approach might consist in public–private alliances. In protected spaces, employees of private companies and representatives of public institutions could collaborate. Public officials would oversee a private company's commitment to suicide prevention and ensure compliance with ethical principles, while readily accepting a private company's engagement in the public good of suicide prevention. Considering now that a limiting condition for the legitimate use of AI for suicide prevention was that it interferes as little as possible with human autonomy, this leads to the conclusion that private companies should be allowed to operate such algorithms only under a set of conditions to be further elaborated and implemented, especially in the domain of confidentiality.

8 All or Nothing?

Given this reasoning, Facebook could simply stop its involvement. An interface between the company's data treasure trove and public agencies could be highly unattractive from Facebook's point of view. If there is no obligation to help and if the commitment is not appreciated, then the company could simply decide against it. Would not this be too high a price to pay if suicides could be prevented? Should not we put the concerns aside and let Facebook do its thing?

One last time, looking at Dave and his company proves helpful: Above it was argued that Dave might be obligated to cooperate under certain conditions. While he has no obligation to become active himself, he does have an obligation to make this help possible. This could also be the case with Facebook. Of course, careful consideration is required here. In particular, it would (again) be necessary to know how well the

algorithms work, how many cases of suicide can really be prevented, and how an intrusion into the company's secrets would have to be evaluated. Too little is known about all this, and Facebook has been criticized for not being willing to reveal details of its suicide detection algorithms (Barnett & Torous, 2019; D'Hotman & Loh, 2020). But it does not have to be an all-or-nothing decision between "Facebook does it itself" and "Nobody does it." On the contrary, well-balanced and nuanced solutions are needed for an area as sensitive as the public treatment of suicidal intent.

A well-balanced approach would also have to include considerations of the proportionality of measures. In this regard, a public–private alliance may prove to be too deep an intrusion into a private company. Externally enforced rules safeguarding confidentiality in combination with an independent evaluation may serve the same goals and prove more proportionate. This requires a framework which entails efficient measures to monitor the implementation of ethical principles for the use of AI by private companies. In this context, lessons can be learned from similar efforts that aim to support the overall goal of "responsible AI" by promoting the corporate implementation of ethical principles. However, the discussion about best practice(s) is far from settled, as de Laat recently observed "(self-)regulation of AI is just starting to crystallize" (de Laat, 2021, p. 1187).

According to our analysis, public bodies are not only entitled to act, but they are on their part obligated to create the necessary conditions for an ethically justifiable use of AI on social media data with the goal of suicide prevention (cf. for the need of public regulation also Coppersmith et al., 2018; D'Hotman et al., 2021; Marks, 2019). This may include private companies participating in suicide prevention efforts when these efforts are embedded in a framework that adequately protects autonomy. As we have shown, this contains important provisions on confidentiality issues in particular. In addition to the question of whether and to what extent private companies are entitled to engage in suicide prevention, the discourse should therefore also address the scope of the obligations of public bodies in establishing and using AI for suicide prevention. It should be noted that Facebook is currently not allowed to deploy its suicide prevention algorithms in the EU because they conflict with data protection regulations (Murphy, 2017). This amounts to an all-or-nothing position on the public side. Given the excessively negative impact of suicide worldwide, public inertia seems unacceptable, both in terms of its own commitment to suicide prevention as well as in terms of an only partially regulated engagement of private companies in this area.

One could reply that Facebook's broad use of AI for suicide prevention is currently unique. Admittedly, we have focused on this particular case. However, it is highly likely that other private companies will follow incentives to use AI for suicide prevention. Public action is therefore also needed anticipating further initiatives in the social media sector and related initiatives of other private companies in other domains (Gomes de Andrade et al., 2018). Even more, states could actively support private companies' initiatives which meet to the conditions sketched above by providing funding or making anonymized health data available to improve research (D'Hotman & Loh, 2020). Overall, instead of all-or-nothing approaches, compromises are needed on both sides: public authorities must create appropriate framework conditions and private companies must adhere to these.

9 Conclusion: Who Should Use AI to Prevent Suicide, if Anybody, and by What Means

In this paper, we addressed the question of whether AI should be used for suicide prevention. Our answer is yes. The principle of beneficence suggests that people with suicidal ideation should be helped. However, there is no obligation for private companies to provide this type of help. In any case, the comparison with established moral practices speaks for this point of view. What is more, specific constraints must be taken into account when applying the principle of beneficence. In fact, it may conflict with the principle of autonomy, so that ethically convincing trade-offs need to be found. The voluntary development and use of AI for suicide prevention by private companies may look like meritorious acts at first glance, but can prove to be ethically problematic on closer inspection. The crucial factor, therefore, is who provides help for persons at risk of committing suicide and under what conditions. We have argued that the effectiveness of AI algorithms must be independently evaluated and that confidentiality must be guaranteed by all means. Only then are the interferences with people's autonomy justified and outweighed by the possible protection of life. Private companies like Facebook can play an important role in suicide prevention if they comply with these rules. At the same time, public bodies have an obligation to create appropriate framework conditions. Working together, public and private institutions can make an important contribution to combating suicide and, in this way, put the principle of beneficence into practice.

Acknowledgements We would like to thank the participants of the colloquium of the Neuroethics group in Jülich as well as the anonymous reviewer for important and instructive comments.

Author Contribution All authors listed have made substantial, direct, and intellectual contributions to the work and approved it for publication.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication. Not applicable.

Competing Interests The authors declare no competing interests.

References

Aquinas, T. (1947). *The summa theologiae of St. Thomas Aquinas* (Fathers of the English Dominican Province, Trans.). Second and revised ed. Benziger Brothers.

Aristotle (1925). *Nicomachean ethics* (W. D. Ross, Trans.). Clarendon Press.

Augustine (1972). *The city of God* (H. Bettenson, Trans.). Penguin Books.

Barnett, I., & Torous, J. (2019). Ethics, transparency, and public health at the intersection of innovation and Facebook's suicide prevention efforts. *Annals of Internal Medicine, 170*(8), 565–566.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health, 17*(16), 5929. https://doi.org/10.3390/ijerph17165929

Camus, A. (2005). *The myth of Sisyphus* (J. O'Brien, Trans.). Penguin.

Celedonia, K. L., Corrales Compagnucci, M., Minssen, T., & Lowery Wilson, M. (2021). Legal, ethical, and wider implications of suicide risk detection systems in social media platforms. *Journal of Law and the Biosciences, 8*(1), 1–11. https://doi.org/10.1093/jlb/lsab021

Cicero, M. T. (1931). *De finibus bonorum et malorum* (H. Rackham, Trans.). W. Heinemann.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomed Inform Insights, 10*, 1–11. https://doi.org/10.1177/1178222618792860

D'Hotman, D., Loh, E., & Savulescu, J. (2021). AI-enabled suicide prediction tools: Ethical considerations for medical leaders. *BMJ Leader, 5*(2), 102–107. https://doi.org/10.1136/leader-2020-000275

de Laat, P. B. (2021). Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy & Technology, 34*(4), 1135–1193. https://doi.org/10.1007/s13347-021-00474-3

D'Hotman, D., & Loh, E. (2020). AI enabled suicide prediction tools: A qualitative narrative review. *BMJ Health Care Inform, 27*(3), e100175. https://doi.org/10.1136/bmjhci-2020-100175

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gomes de Andrade, N. N., Pawson, D., Muriello, D., Donahue, L., & Guadagno, J. (2018). Ethics and artificial intelligence: Suicide prevention on Facebook. *Philosophy & Technology, 31*(4), 669–684. https://doi.org/10.1007/s13347-018-0336-0

Helm, B. (2021). Friendship. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Fall 2021 ed.). https://plato.stanford.edu/archives/fall2021/entries/friendship/

Hume, D. (2005). *On suicide*. Penguin.

Laacke, S., Mueller, R., Schomerus, G., & Salloch, S. (2021). Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. *Am J Bioeth, 21*(7), 4–20. https://doi.org/10.1080/15265161.2020.1863515

Lutz, P. E., Mechawar, N., & Turecki, G. (2017). Neuropathology of suicide: Recent findings and future directions. *Molecular Psychiatry, 22*(10), 1395–1412. https://doi.org/10.1038/mp.2017.141

Marks, M. (2019). Artificial intelligence-based suicide prediction. *Yale JL & Tech*(21), 98–121. https://ssrn.com/abstract=3324874

Murphy, M. (2017). *EU data laws block Facebook's suicide prevention tool*. The telegraph.

O'Dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L., & Christensen, H. (2017). A linguistic analysis of suicide-related Twitter posts. *Crisis, 38*(5), 319–329. https://doi.org/10.1027/0227-5910/a000443

Plato (1951). *Phaedo* (F. J. Church, Trans.). The Liberal Arts Press.

Pourmand, A., Roberson, J., Caggiula, A., Monsalve, N., Rahimi, M., & Torres-Llenza, V. (2019). Social media and suicide: A review of technology-based epidemiology and risk assessment. *Telemedicine Journal and E-Health, 25*(10), 880–888. https://doi.org/10.1089/tmj.2018.0203

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digit Med, 3*, 78. https://doi.org/10.1038/s41746-020-0287-6

Sedgwick, R., Epstein, S., Dutta, R., & Ougrin, D. (2019). Social media, internet use and suicide attempts in adolescents. *Current Opinion in Psychiatry, 32*(6), 534–541. https://doi.org/10.1097/yco.0000000000000547

Singer, N. (2018). Screening for suicide risk, Facebook takes on tricky public health role. *New York Times* (2018, December 31). https://www.nytimes.com/2018/12/31/technology/facebook-suicide-screening-algorithm.html Accessed 11 Jul 2022

Sueki, H. (2015). The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of Affective Disorders, 170*, 155–160. https://doi.org/10.1016/j.jad.2014.08.047

Wittgenstein, L. (1984). *Notebooks 1914–1916* (G. E. M. Anscombe, Trans.). Second ed. University of Chicago Press.

World Health Organisation. (2021). *Suicide worldwide in 2019: Global health estimates*. World Health Organization.