



Merging public health and automated approaches to address online hate speech

Tina Nguyen¹

Received: 27 December 2022 / Accepted: 26 March 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

The COVID-19 pandemic sparked a rise in misinformation from various media sources, which contributed to the heightened severity of hate speech. The upsurge of hate speech online has devastatingly translated to real-life hate crimes, which saw an increase of 32% in 2020 in the United States alone (U.S. Department of Justice 2022). In this paper, I explore the current effects of hate speech and why hate speech should be widely recognized as a public health issue. I also discuss current artificial intelligence (AI) and machine learning (ML) strategies to mitigate hate speech along with the ethical concerns with using these technologies. Future considerations to improve AI/ML are also examined. Through analyzing these two contrasting methodologies (public health versus AI/ML), I argue that these two approaches applied by themselves are not efficient or sustainable. Therefore, I propose a third approach that combines both AI/ML and public health. With this proposed approach, the reactive side of AI/ML and the preventative nature of public health measures are united to develop an effective manner of addressing hate speech.

Keywords Artificial intelligence · Public health · Machine learning · Hate speech · Hate crimes · Cyberbullying

1 Introduction

During the start of the COVID-19 pandemic, the world faced what is known as an “infodemic” or an overload of information, whether false or not, in the time of a disease outbreak [1]. With the capability of mobile technology to easily access information instantaneously, it became increasingly difficult to control the narrative surrounding the disease. Thus, misinformation surrounding COVID-19 alongside the uncertainty of the pandemic and subsequent national lockdown led to online hate speech targeting China and people of Chinese descent [2]. Fan et al.’s study, which analyzed hate speech on Twitter, reported an estimated 25,000 hate speech-related tweets within the United States (U.S.) alone [2]. By mid-March of 2020, this online hate translated to offline violent acts directed towards the Asian American and Pacific Islander (AAPI) community [3]. In 2020, the U.S. Department of Justice reported a 77% surge in Anti-Asian hate crimes compared to the previous year [4]. Unfortunately,

the AAPI community is not the only group that has been subjected to receiving online hate speech, nor has it been the only victims of hate crimes recently or in the last several years. Preceding Anti-Asian hate, other minority groups such as Black, Jewish, Latine, Muslim, Indigenous populations, LGBTQ community, and women have also received hate speech. Additionally, 2020 also saw a 32% increase in hate crimes [4]. While not all instances of hate speech are correlated to hate crimes, hate speech does influence how minorities are perceived by society [5]. In turn, one of the consequences of these misperceptions is the progression from hate speech to hate crimes.

While hate crimes are considered a public health issue in the U.S., hate speech are not [6]. Recognizing hate crimes as a public health issue brings awareness and invokes action, such as new policies being formed or allocating resources to help victims, whereas not formally acknowledging hate speech as a public health issue downplays the act and delegitimizes the effects on victims. Similarly, cyberbullying in adolescents has also been identified as a public health issue [7, 8]. By recognizing cyberbullying as a public health issue, the U.S. federal government enacted initiatives to mitigate the effects of cyberbullying, such as guidance on school policies and

✉ Tina Nguyen
tinguye1@utmb.edu

¹ University of Texas Medical Branch, 301 University Blvd, Galveston, TX 77555, USA

reporting systems [9]. While both hate crimes and cyberbullying are related to hate speech in different ways, these societal concerns have mechanisms to deal with these concerns because of their status as public health issues. Classifying hate speech as a public health issue serves as one of the approaches to addressing hate speech, whereas several studies have shown the effectiveness of artificial intelligence (AI) and machine learning (ML) in identifying as well as stopping hate speech [10–12]. However, there are also ethical consequences that come with utilizing AI/ML to curb hate speech, such as unfairness in silencing users [13]. Given the role technology has in disseminating hate speech while also being a possible solution to this issue, it is highly relevant now more than ever to discuss the effects of this long-term societal concern. Therefore, this paper seeks to: (1) understand how hate speech can be a public health issue, (2) explore the benefits of utilizing AI/ML as well as the harms from these technologies, (3) discuss future considerations for AI/ML regarding hate speech, and, finally, (4) examine a proposed approach that combines both public health and AI/ML.

2 Hate speech: a potential public health issue

Despite the spike in hate speech during the onset of the pandemic, hate speech has been an ongoing issue. The exact origins are unknown. However, in 1989, the U.S. officially named this toxic language towards outgroup populations as “hate speech” [14]. According to the United Nations (UN), hate speech are

any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor [15].

Meanwhile, the website Hatebase, a database of derogatory terms covering 98 languages, identified over a million instances of hateful words being used online throughout the world [16]. With the emergence of social media, spreading hate speech to a larger audience has turned into a simplified process. Targeted harassment towards minority groups has become exacerbated with social media sites, which allows users to spew hate messages almost instantaneously [17]. In this section, I discuss the harms from hate speech, how public health would approach hate speech, and examine related public health issues to further develop hate speech as a public health issue.

2.1 Harms from hate speech

The famous quote “sticks and stones may break my bones, but words will never hurt me” is often used by victims to show that they are physically and mentally stronger than the damaging words being thrown at them. While the intention behind this quote is to empower the victim by downplaying the weight of harmful language, unfortunately, this is not always true. Hate speech can cause a variety of harms towards the victim, groups of people, and society. These include mental, emotional, social, and physical harms (through escalation to hate crimes) [18]. Some individuals might argue that while they do not like witnessing hate speech towards outgroup members/minority groups, victims of hate speech “[...] should just learn to live with it” as hate speech are just words [19]. As ingroup members, it is easy to ignore hate speech as it does not personally affect these individuals, but as outgroup members being subjected to hate speech affects them on a long-term basis [20]. Negative emotions, such as “anger, shame, and fear”, arise with hate speech, which can sometimes cause internalized hate [18]. Additionally, victims of hate speech face mental health issues, such as posttraumatic stress disorder (PTSD), anxiety, suicidal thoughts, and depression [18]. Being on the receiving end of hate speech can detrimentally affect how a person views their own self-worth and dignity [20]. Thus, unlike ingroup members, outgroup members do not have the privilege to simply disregard hate speech as this directly impacts them in a multitude of ways.

Meanwhile, the shift from online hate speech to offline hate crimes causes not only physical but also societal harm. Several scholars in the field claim that there is a correlation between online hate speech and offline hate crimes [21, 22]. Particularly, Cahill et al.’s study found that hate speech tweets in addition to other factors, such as geographic area, were risk factors for hate crimes against both racial and religious minorities [23]. While the pipeline from causing emotional and mental harms due to spreading online hate speech to physically harming victims through a hate crime might seem rather drastic, this transition is often not an immediate one. Instead, it slowly builds over time. Repeated exposure to hate speech can weaken its severity, thus inducing “desensitization” from non-affected individuals to such language [24]. Similarly, desensitization causes ingroup members to become less empathetic or unsympathetic towards outgroup communities [25]. Constant subjection to online hate speech can also establish a new “normalized” view of minority groups [25]. This skewed perspective on minority groups causes societal harm as outgroup members are now seen by society as inferior compared to the ingroup/majority groups [25]. An

accumulation of apathy towards hate speech and negative, stereotypical perspectives on minorities can develop into radicalization, which in the offender's mind justifies the hate and consequential violence against outgroup members [26]. For example, the Tree of Life Synagogue shooting in Pittsburgh the perpetrator, who already had antisemitic views, became further radicalized by frequently visiting and interacting with like-minded individuals on an unregulated social media site filled with hate speech against Jews [25, 27]. This tragedy represents two different messages. For society, it shows the harms of hate speech and radicalization. Meanwhile, for individuals with the same views as the perpetrator's, this illustrates "acceptable" behavior towards an outgroup population and might even radicalize others exposed to this tragedy. Therefore, hate speech are more than just words to be ignored. The many harms from this verbal attack are detrimental to the victims and should not be their burden to face alone.

2.2 Public health's approach to addressing hate speech

According to the Institute of Medicine (IOM), "Public health is what we as a society do collectively to assure the conditions in which people can be healthy" [28]. Similarly, several definitions of public health embrace a broad scope of what society's role is in improving or promoting the population's health by considering other factors that might affect their health, such as societal or cultural nuances [29–32]. Traditional health problems that are addressed by public health efforts include increasing vaccination rates, having screening programs, and providing educational resources for prevention [32]. Meanwhile, the broad view of public health covers a range of issues, such as homelessness or domestic violence, in addition to the traditional health problems [31, 32]. However, some suggest that there should be a narrow view or definition of public health instead of a broad scope, as the main objectives of public health are lost once extraneous issues are considered [31, 32]. While I agree that not every issue needs to be classified as a public health issue, narrowing the scope of what counts as a public health issue diminishes the legitimacy of the harms faced by typically vulnerable communities. Therefore, I believe that public health should continue to adopt a broad scope definition of public health to be inclusive to the health needs of all populations and to foster a healthy environment for all. Detels and Tan also believe that public health should cover a broad scope of issues, to meet the World Health Organization's (WHO) mission of "health for all", rather than focusing on just disease prevention [33].

2.2.1 Related public health issues and their strategic actions

If we apply the broad view to public health, then hate speech would fall under the category of hate-motivated behavior [34]. According to Shultz et al., "The direct consequences of hate—including violence, discrimination, and marginalization of outgroups—are associated with poor health" [34]. Whether mental, physical, emotional, or social harms, hate-motivated behavior negatively affects the livelihood of outgroup members—some of whom are already dealing with other social determinants of health [35]. Hate-motivated behavior goes against public health's aim of enhancing the health of communities [29]. Minority populations cannot have a healthy lifestyle if they are actively targeted for hate speech. Similarly, witnessing such violence can affect the well-being of nonparticipants and create an overall unsafe environment. Examining related public health issues, such as hate crimes and cyberbullying in adolescents, will further supplement my argument of why hate speech should be a public health issue.

The acknowledgment of hate crimes as a public health issue by the American College of Physicians essentially placed pressure on other organizations to do the same, which allowed for more discussions and strategic actions to be done [34, 36]. These strategic actions include creating a task force for collecting racial inequity data, engaging with communities and community leaders, and developing policies that address racial inequities [36]. Meanwhile, state and federal governments have also taken actions to prevent hate crimes, such as New York creating the Office for the Prevention of Hate Crimes (OPHC) and enacting several hate crime prevention acts [37]. Unfortunately, due to the variation in states' hate crime laws, hate crimes in the U.S. were at an all-time high in 2021 [38]. Collective and standardized policies that address the shortcomings of current policies, such as inconsistent data collection and biases, should help in mitigating hate crimes [38]. Likewise, cyberbullying in adolescents is another related public health issue that benefitted from this classification. Strategic actions for cyberbullying in adolescents include empathy training, parental control of their children's time online, and school programs addressing cyberbullying [7, 8]. In Gaffney et al.'s study, they reported an estimated 21 to 25% decrease in bullying incidences in the U.S., Norway, and Italy with the implementation of bullying prevention programs [39]. Although both hate crimes and cyberbullying in adolescents continue to be persistent issues, the incremental progress due to their status as public health issues has helped in fostering collective action. Structural changes cannot happen if no one is aware of or actively discussing the issue. Therefore, just as hate crimes and cyberbullying are categorized as public health issues, hate speech should also be recognized as a public health

issue. Hate speech are an enduring concern that warrants the public health issue title. Not only does hate speech affect outgroup populations, but it also has the possibility of turning into a hate crime. One of the priorities for the strategic actions for curbing hate speech will be the de-escalation of desensitization and radicalization. The Cure Violence (CV) model by Dr. Gary Slutkin seeks to prevent violent behavior through “[...] interrupting transmission directly, identifying and changing the thinking of potential transmitters (i.e., those at highest risk of perpetrating violence), and changing group norms regarding violence” [40]. Adopting the CV model’s objectives, other proposed strategic actions could potentially include policies denouncing this language, having cultural diversity classes or resources available to the public, and investing in resources for victims.

3 The promise and pitfalls of using AI/ML to combat hate speech

As social media platforms increase daily user numbers and their functionalities, hate speech can come in various forms besides spoken words. Whether typed, in a video format, or through images, social media can spread hate speech rapidly and expansively. According to Hawdon et al., 53% of U.S. citizens (ages 15–30 years) have come across hate speech at some point in their time online [41]. Thus, various research studies have focused on identifying and mitigating online hate speech using AI/ML. Popular social media sites have even released statements denouncing hate speech and what actions they are taking to eradicate hate speech on their platforms. For example, Meta, more commonly known as Facebook, claimed that its algorithm identified and deleted 97% of hate speech-related posts [42]. Meanwhile, Twitter has taken a more cautious approach by broadly addressing hate speech as “hateful conduct” and enacting specific actions, such as removing tweets or suspending users, per their hateful conduct policy [43].

Current strategies that most social media companies utilize include suspending accounts, deleting inappropriate comments, and terminating accounts. Even though these actions can stop hate speech to a certain extent, they are only temporary solutions. For example, suspending accounts is not sustainable as some accounts can either be unsuspended if the user appeals or they could create another account, which defeats the purpose of the enacted policies [21]. Additionally, before implementing their algorithm, Facebook utilized human moderators to evaluate flagged content, which is too much for a single person to realistically manage given the large amount of big data accumulated from Facebook posts [12]. Even before the content moderation stage, someone must flag the inappropriate content for review. As a result, this individual is being exposed to hate speech.

This current system, which Twitter and Reddit are presently employing, places the onus on its users to flag hateful content they encounter [44]. The burden of essentially being an online “vigilante” of taking down hate content should not be on the users, but instead on the companies. Ideally, minimizing hate speech exposure for both outgroup and ingroup populations should be a priority for social media companies.

Research on innovative AI/ML that can mitigate hate speech is abundant, with most studies focusing on different methods to combat hate speech, such as detecting or decreasing instances of hate speech [10–12]. Instant detection of hate speech from AI/ML can save content moderators’ time and prevents users from interacting with hate speech. The methods for hate speech detection have seemingly shifted from the bag-of-words (BoW) approach, which trains algorithms to focus on specific words to detect if the post uses hate speech, to algorithms using natural language processing (NLP) filtering tools and sentiment analysis to identify hate speech [10, 11, 45]. Although these efforts from researchers and social media platforms are meant to mitigate hate speech, there are still issues with AI/ML usage, such as bias in what gets flagged as hate speech. In the following sections, I will explore the concerns of using AI/ML and future considerations for this technology.

3.1 Concerns with utilizing AI/ML to mitigate hate speech

While there are benefits to using AI/ML to control hate speech, there are still workarounds to spreading hate speech. Besides malicious users actively creating new accounts to spew hate speech onto social media sites, these individuals have also found ways to bypass the algorithms [14]. Purposely misspelling words or using “leetspeak” (e.g., “134rn” which equates to “learn”) are just some of the methods bad actors employ to avoid hate speech detection [14]. Another example is deliberately putting spaces in between each letter to spell derogatory words [14]. These manipulative techniques confuse the algorithm and allow for hate speech to be undetected. Relatedly, offenders can word their posts in a manner that expresses implicit hate speech, which algorithms also have difficulty recognizing [46]. Explicit hate speech occurs when the hateful intent behind the words is obvious, whereas implicit hate speech are described as “coded” language and thus become complicated for AI/ML to decipher the true nature of the words [46]. Phrases like “send them home” or “get them out” are considered as implicit hate speech due to the vagueness of who is targeted and, more significantly, the context behind the phrases [10]. An algorithm cannot fully grasp the intention behind those phrases as it is trained to pick up on certain words or direct phrases.

In addition to user workarounds comes the problem of AI/ML not understanding the nuance of different dialects, causing biased outcomes [45, 47]. Dias Oliva et al.'s study found that drag queens were more likely to have their tweets flagged as having high levels of toxicity due to their vernacular in comparison to White nationalists [45]. For example, the AI tool "Perspective" categorized an estimated 6% of drag queens' tweets as having 70% or higher toxicity, whereas approximately 3% of White nationalists' tweets had toxicity levels of 70% or higher [45]. Therefore, Perspective is only taking the face value of the word(s), instead of considering the context or intent behind the sentence. Similarly, Perspective AI ranked Black users' toxicity higher than other users [47]. For instance, the tweet "I saw his ass yesterday" had a toxicity score of 95% in comparison to "I saw him yesterday", which only had a 6% toxicity level [47]. As minority groups have reclaimed derogatory terms, such as the f slur or the n-word, or have certain colloquialisms, such as African American Vernacular English (AAVE), it is erroneous for AI/ML to label these tweets as having higher toxicity levels than instances of implied or explicit hate speech [45, 47]. Although these results display AI/ML reinforcing bias, the AI/ML should not be solely blamed for these problematic findings. Algorithms require data for training and are often unable to comprehend if their data are biased. As such, annotators play a significant role in the quality of the training datasets [48, 49]. Waseem's study found that amateur annotators tended to misclassify something as hate speech more often than their expert counterparts [48]. Consequently, the algorithms trained with datasets vetted by expert annotators had increased accuracy in comparison to datasets from amateur annotators [48]. Biased AI/ML, and to an even further extent biased datasets, are detrimental to hate speech detection efforts and may lead to unjustly suspending or silencing accounts based on implicit bias [45].

Furthermore, just as AI/ML can be used to reduce hate speech, it can also incite hate speech. According to Albadi et al., approximately 15% of Twitter accounts are associated with social bots, totalling to a significant 48 million social bot accounts [50]. Social bots are algorithms that imitate users' activity on social media often with an ulterior motive [51]. While social bots are generally benign, in recent years, they have been commonly associated with unvirtuous purposes [51]. For example, during the 2016 U.S. election, it was discovered that social bots were used to interfere with the election results [52]. Social bots spread misinformation that could have influenced some users in their decision-making [53]. As evidenced by the U.S. election, social bots can affect how human users perceive specific topics. Likewise, social bots can develop a false narrative that can harm outgroup populations and disseminate hate speech [50, 54]. Social bots act as a collective by interacting with each other to make other users assume the bots' message is commonly

shared [53]. The bots' ability to normalize misinformation relates to hate speech exposure and the subsequent desensitization of individuals. Researchers found that bots were responsible for spreading conspiracy theories during the early stages of the pandemic, such as COVID-19 coming from a lab in Wuhan and other QAnon-related theories [54]. Bots sharing and boosting such conspiracy theories at a high level can contribute to some users believing these lies as the truth; thus, warping their perception and placing the blame on the AAPI community. The misinformation surrounding COVID-19 origins led to hate speech and, consequently, hate crimes [2, 3]. In addition to misinformation, bots can quickly spread and invoke hate speech. Uyheng and Carley concluded that high bot activity led to increased hate in ingroup populations, further proving social bots' influence [52]. Users who unknowingly follow or interact with social bots are stuck in an echo chamber and can potentially become radicalized [26, 52]. However, the arguments here are not meant to completely dissolve accountability from those who contribute to hate speech online. While bots can easily spread hate speech and are responsible for misinformation, they serve merely as controversy instigators. As evidenced by Albadi et al.'s study, bots created approximately 23% of hate speech tweets, whereas 77% of hate speech tweets were from human users [50].

3.2 Future considerations for AI/ML

While current research is focused on AI/ML recognizing hate speech, the future goal for AI/ML is automated hate speech detection and deletion. By applying this methodology, AI/ML would prevent victims from seeing hate speech, thus precluding desensitization and radicalization. However, with automated hate speech identification and subsequent deletion, concerns about bias and violation of free speech arise [12, 55, 56]. Concerning AI/ML not being able to distinguish between hate speech and the dialect of other communities, automated detection and deletion can have the same effects of silencing outgroups based on the algorithm's assumption that they are stating hate speech. According to Armijo, while algorithms excel at what they are programmed to do, they lack the ability to explain their reasoning [12]. As such, users who believe they were wrongly suspended or were silenced will not understand why these actions happened to them, because the algorithm cannot explain the context behind their decision [12]. Humans will always have a role in moderating hate speech because of the explainability issue [12, 55]. Similarly, some might argue that automated hate speech detection and especially deletion infringes on free speech rights [12, 57, 58]. A comparable argument to silencing outgroups could be made for those who disseminate hate speech, where they believe they are protected under the First Amendment, and thus deleting or

suspending their account is unfairly breaching their rights [57]. Again, the deletion aspect, and to a further extent the subsequent suspension or termination of an account, greatly hinders this process from being fully automated.

Therefore, research interests have also been dedicated to utilizing AI/ML to stop further instances of hate speech through the diffusion of counter-speech [57–59]. Counter-speech is a response from another user that calls out the offensiveness of hate speech and is meant to prevent the original poster from doing it again [59]. However, counter-speech can be expressed through several methods, including “humor, warning of consequences, and inducing empathy” [58]. An example of an empathy-induced counter-speech message is “remember that those you care about can see this post too...” [58]. This message serves as a reminder to the perpetrator that their hateful behavior is publicly available for others to see and invokes a sense of shame as they might not want these comments to be seen by people they know. The goal of counter-speech is not always to directly call out the perpetrator for hate speech, but instead supplying a surplus of content that distracts from hate speech posts [57]. According to Garland et al., “After the emergence of the organized counter group Reconquista Internet (RI) in the late spring of 2018, the relative frequency of counter-speech increased while that of hate speech decreased” [60]. This real-life example from Germany shows that coordinated efforts to spread counter-speech can prevent hate speech. While counter-speech does not prevent victims from seeing every instance of hate speech, it serves as an effective and non-violating option in combatting hate speech [57]. Unlike automated hate speech detection and deletion, counter-speech does not violate free speech, because it does not silence users [58]. As such, future studies regarding counter-speech aim to understand which type of counter-speech is the most useful for each group [57]. Meanwhile, Cypris et al. suggest an automated counter-speech algorithm that can respond to hate speech with either previously drafted counter-speech messages or specifically generated content [56]. In addition to an automated counter-speech algorithm, manipulation of social bots to spread counter-speech instead of hate speech or misinformation has also been proposed [61]. However, in discussing counter-speech and future applications of it, we must also recognize that without proper hate speech detection tools robust counter-speech would be difficult to accomplish.

Another crucial consideration for further research of AI/ML and hate speech is having larger and consistent datasets [14]. There are limited (in quantity and quality) datasets that researchers have been using in their studies, which caused difficulties in testing and developing the functionalities of AI/ML [14, 55]. Data quality depends on the annotators, and the difference in expert and amateur annotators affect algorithmic accuracy [14, 48, 55]. Likewise, the quantity

of datasets has been scarce, which has led to overfitting in some studies [55, 62]. As such, larger and diverse datasets, including non-English datasets, are needed to train future AI/ML efforts [50].

4 Proposed public health and AI/ML approach

Of the two approaches discussed in this paper, neither one by itself can completely eradicate hate speech. Some might argue that public health approaches in general are slow-moving processes, especially in terms of enacting policies [28]. On the other side of the argument, others might insist that AI/ML is not sustainable as simply silencing perpetrators of hate speech does not stop the overall problem in the long run. A recent example of this is the restrictions on Ye’s, also formerly known as Kanye West, Twitter and Instagram accounts [63]. While the social media companies halted the hate speech posts from Ye, they did not prevent further instances of hate speech as an antisemitic group in Los Angeles were seen displaying an antisemitic sign in support of Ye [64]. Similarly, Ye’s Twitter account was shortly reinstated by Twitter even before new leadership by Elon Musk [65]. As such, silencing users, especially celebrities or influencers, whose actions or words can have major effects on the masses, is not a viable solution.

Therefore, I believe that a combination of public health and AI/ML should be done to develop a sustainable response to combatting hate speech. However, with the goals of both public health and AI/ML contrasting from one another, this serves as a barrier to creating a cohesive approach. Public health’s goals are to prevent, whereas with AI/ML, it is to react [12, 40, 44, 66]. These two conflicting goals cause friction between what a proposed approach would look like. For a while, public health was mostly applying AI/ML for surveillance or detection of disease progressions [67, 68]. For example, the early detection of the Zika virus was due to big data analytics using AI/ML [68, 69]. From the information gathered by AI/ML, public health officials and governments have been able to develop collective actions, such as educating the public on measures to prevent illnesses or providing access to interventions [31, 68, 70].

Given the upsurge of the infodemic, the focus of AI/ML in public health has gradually transitioned from epidemiology to “infodemiology”, where data analytics of social media and other internet data sources are used to guide public health officials in developing strategic actions [71, 72]. Infodemiologists use social listening, a method often used in the business industry to track their brand, to monitor the discussion surrounding a topic of interest [73, 74]. An example of this is the WHO’s Early AI-supported Response with Social Listening (EARS), an AI that follows the conversation about

COVID-19 resulting in early detection of misinformation [75]. The initial EARS AI/ML, developed “with proprietary technology from Citibeats”, was trained with social media data from 20 countries, including languages such as English, Spanish, French, and Portuguese, and used semi-supervised learning to generate its outcomes [74, 76]. According to Purnat et al., “Data [were] categorized automatically into 40 categories [about COVID-19 conversations], with human quality controls, as per the WHO-developed COVID-19 public health social listening taxonomy” [74]. The social listening taxonomy included topics such as “the cause of the virus”, who was at risk for the illness, what interventions (e.g., testing, healthcare equipment) are available, and tracking of misinformation [71]. The pilot study has since increased to 30 countries and includes an additional 26 categories dedicated to conversations surrounding the COVID-19 vaccine [77]. Using EARS to monitor conversations and address misinformation in an efficient manner can lessen the effects of the infodemic as well as the stigmatization of the AAPI community.

4.1 AI/ML for social listening of hate speech and public health interventions

As previously mentioned, automated deletion interferes with free speech rights and is akin to silencing outgroups [12, 57, 58]. Therefore, in proposing a joint process, I argue that AI/ML should only be used for automated hate speech detection and data analytics. These outcomes from AI/ML will then be used to inform the public health agenda. Using the EARS project as a reference, my proposed approach involves using an algorithm for social listening to monitor hate speech which will then guide the public health efforts. Through the results of automated hate speech detection and data analytics, we can track areas where hate speech are prevalent and find root causes concerning the misconceptions about outgroups. Unfortunately, however, this might be difficult to quantify given the continuous growth of conversations surrounding hate speech. Meanwhile with EARS, the conversations surrounding COVID were easier to track and quantify into specific categories, because eventually the topic of COVID came to a standstill, where nothing ground-breaking was being contributed to the discussion. As such, researchers were able to discern categories and subcategories much quicker than with the ever-expanding topic of hate speech. Additionally, there are many groups that already are or could be considered as an outgroup, which further complicates what categories and subcategories should be tracked. For the purposes of illustrating my proposal, an initial AI/ML for social listening of hate speech could include the categories of race/ethnicity (e.g., hate speech against Black, Asian, and Latine), religion (e.g., islamophobia, antisemitism), sexual orientation, and gender. As this AI/ML gains traction and

adequate resources, more categories/subcategories could be added for future iterations. Similarly, expert annotators are needed to help determine the appropriate category for each post and will act as quality control—ensuring that there are consistencies in the automatic detection of hate speech. From this information, I believe that the following public health efforts will support the mitigation of hate speech: correcting misinformation through educational resources, working with different groups to develop prevention programs, and developing policies to address non-tolerance of harmful behavior. Some future considerations for this proposed approach are to expand the languages studied and to include other formats besides textual analysis [78]. Most studies for social listening of any sort primarily focus on English, which dismisses the fact that hate speech could be in any language. Meanwhile, Matamoros-Fernandez and Farkas’ literature review showed that most studies about hate speech mainly concentrated on Twitter (despite other social media sites, such as Facebook and WhatsApp, having more users) and text-based posts instead of other formats, such as photos or videos [78].

A barrier to my suggested approach is arguments against paternalism. Paternalism is often cited as the reason why some resist public health efforts despite the good intentions as they believe these interventions infringe on their autonomous right to decide for themselves what their wants and needs are [30]. Given that “[t]he providers of public health interventions often are governments, rather than private practitioners”, there is an added opposition from some users to regulate social media companies due to the fears of content moderation as well as silencing users [29, 79]. Currently, social media companies are private corporations and function on their own terms often without government regulation in most countries [80, 81]. Even with public health policies enacted by the government, this cannot be enforced onto the social media companies as they have “[...] their own rules for acceptable user behaviour” [82]. However, Abroms states that

[o]ne exception to the voluntary nature of the guidelines is a new, first-of-its-kind law in Germany, which mandates that hate speech, fake news, and illegal material must be taken down within 24 hours of being reported, or companies could be fined millions of dollars by the government [80].

This law, while most likely effective, is a tad unforgiving. Ideally, regulated social media websites or the Internet would protect affected communities. Similarly, there have been debates on reforming Section 230 of the Communications Decency Act, which would hold social media companies liable for the content they host on their website [79, 83]. Both regulation and reforming Section 230 are direct interference from the government and considered as paternalism

due to these actions, with the intention of preventing hate speech, affecting the users, who might not have the same, normative views towards outgroups. Morley et al. argue that “[m]ore coercive forms of information control are perceived to be neither necessary nor proportionate” [84]. Again, the goal is to prevent hate speech and not silence users. Despite the benefits of government intercession, I believe that regulation of social media companies and reforming Section 230 in the U.S. is less likely to happen. Therefore, instead of focusing on regulation, which will be different for each country (and state-wide in the U.S.) anyway, and reforming Section 230, I recommend governments to redirect their attention to implementing sustainable interventions. Funds towards resources that help outgroups and education that discusses diversity, equity, and inclusion (DEI) should be advocated for by the government. By focusing on strategic actions to stop hate speech, instead of enforcing “coercive” actions, the goal of eradicating hate speech not only online but in society can be achieved.

5 Conclusion

Unfortunately, the issue of hate speech will not cease to exist. This hateful mentality comes from years of desensitization, and with groups being more covert with their messages, it has become increasingly difficult to stop these false narratives surrounding outgroup populations. Relatedly, hate crimes and cyberbullying are considered as public health issues, which has afforded more public awareness and resources for causes fighting against both issues. Therefore, the classification of hate speech as a public health concern can amplify the severity of the harms stemming from hate speech. Several studies on hate speech detection by AI/ML have shown early promise in its functionality [10, 11, 45]. However, ethical consequences to AI/ML usage, such as unfairness, arise in preventing hate speech. Nevertheless, given AI/ML’s capabilities to analyze big data in social media and lessen the emotional and physical burden on human moderators, it could potentially be mandatory in the future to utilize these tools to control hate speech [12]. Furthermore, this paper discussed a proposed approach to addressing hate speech. While public health and AI/ML have different goals, I believe that uniting the reactive side of AI/ML with the preventative side of public health will result in a sustainable and efficient approach to confronting the problem of hate speech.

Data availability The author did not analyse or generate any datasets, because the work proceeds within a conceptual and normative ethics approach. One can obtain the relevant materials from the references below.

Declarations

Conflict of interest The author declare that they have no conflict of interest.

References

1. World Health Organization, Bradd, S.: Infodemic. https://www.who.int/health-topics/infodemic#tab=tab_1 (2022). Accessed 5 Aug 2022
2. Fan, L., Yu, H., Yin, Z.: Stigmatization in social media: documenting and analyzing hate speech for COVID-19 on twitter. *Assoc. Inf. Sci. Technol.* **57**(1), 1–11 (2020). <https://doi.org/10.1002/pr2.313>
3. Gover, A.R., Harper, S.B., Langton, L.: Anti-Asian hate crime during the COVID-19 pandemic: exploring the reproduction of inequality. *Am. J. Crim. Justice* **45**, 647–667 (2020). <https://doi.org/10.1007/s12103-020-09545-1>
4. U.S. Department of Justice: 2020 FBI hate crime statistics. Last Modified: May 5. <https://www.justice.gov/crs/highlights/2020-hate-crimes-statistics> (2022). Accessed 7 July 2022
5. Weber, M., Viehmann, C., Ziegele, M., Schemer, C.: Online hate does not stay online – how implicit and explicit attitudes mediate the effect of civil Negativity and hate in user comments on prosocial behavior. *Comput. Hum. Behav.* **104**, 1–10 (2020). <https://doi.org/10.1016/j.chb.2019.106192>
6. American College of Physicians: American college of physicians says hate crimes are public health issue. Last Modified August 14. <https://www.acponline.org/acp-newsroom/american-college-of-physicians-says-hate-crimes-are-public-health-issue> (2017). Accessed 19 July 2022
7. Nixon, C.L.: Current perspectives: the impact of cyberbullying on adolescent health. *Adolesc. Health Med. Ther.* **5**, 143–148 (2014). <https://doi.org/10.2147/AHMT.S36456>
8. Zhu, C., Huang, S., Evans, R., Zhang, W.: Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Front. Public Health* **9**, 1–12 (2021). <https://doi.org/10.3389/fpubh.2021.634909>
9. Stopbullying.gov: Prevent cyberbullying. Last Modified November 10. <https://www.stopbullying.gov/cyberbullying/prevention> (2021). Accessed 13 Oct 2022
10. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy Internet* **7**(2), 223–242 (2015)
11. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. *Proc. Fifth Int. Worksh. Natl. Lang. Process. Soc. Media* (2017). <https://doi.org/10.18653/v1/W17-1101>
12. Armijo, E.: Speech regulation by algorithm. *William Mary Bill Rights J.* **30**(2), 1–13 (2021)
13. Carlson, C.R.: Hate speech as a structural phenomenon. *First Amend. Stud.* **54**(2), 217–224 (2020). <https://doi.org/10.1080/21689725.2020.1837649>
14. Grondahl, T., Pajola, L., Juuti, M., Conti, M., Asokan, N.: All you need is love: evading hate speech detection. *Proc. ACM Worksh. Artif. Intell. Secur.* (2018). <https://doi.org/10.48550/arXiv.1808.09115>
15. United Nations: United Nations strategy and plan of action on hate speech. <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf> (2019). Accessed 14 August 2022
16. Hatebase. <https://hatebase.org/> (2022). Accessed 10 July 2022

17. Saurwein, F., Spencer-Smith, C.: Automated trouble: the role of algorithmic selection in harms on social media platforms. *Media Commun.* **9**(4), 222–229 (2021). <https://doi.org/10.17645/mac.v9i4.4062>
18. Cramer, R.J., Fording, R.C., Gerstenfeld, P., Kehn, A., Marsden, J., Deitle, C., King, A., Smart, S., Nobles, M.R.: Hate-motivated behavior: impacts, risk factors, and interventions. *Health Affairs Health Policy Brief* **9**, 1–6 (2020)
19. Waldron, J.: *The Harm in Hate Speech*. Harvard University Press, Cambridge (2014)
20. Gelber, K., McNamara, L.: Evidencing the harms of hate speech. *Soc. Identities* (2015). <https://doi.org/10.1080/13504630.2015.1128810>
21. Saha, K., Chandrasekharan, E., De Choudhury, M.: Prevalence and psychological effects of hateful speech in online college communities. *Proc. ACM Web Sci. Conf.* (2019). <https://doi.org/10.1145/3292522.3326032>
22. Reichelmann, A., Hawdon, J., Costello, M., et al.: Hate knows no boundaries: online hate in six nations. *Deviant Behav* **42**(9), 1100–1108 (2021). <https://doi.org/10.1080/01639625.2020.1722337>
23. Cahill, M., Migacheva, K., Taylor, J., et al.: Understanding online hate speech as a motivator and predictor of hate crime. National Criminal Justice Reference Service. <https://www.ojp.gov/pdffiles1/nij/grants/304532.pdf> (2019). Accessed 2 Nov 2022
24. Soral, W., Bilewicz, M., Winiewski, M.: Exposure to hate speech increases prejudice through desensitization. *Aggress. Behav.* **44**, 136–146 (2017). <https://doi.org/10.1002/ab.21737>
25. Bilewicz, M., Soral, W.: Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Adv. Polit. Psychol.* **41**(1), 3–28 (2020). <https://doi.org/10.1111/pops.12670>
26. Hatzipanagos, R.: How online hate turns into real-life violence. *The Washington Post*. <https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence/> (2018). Accessed 20 July 2022
27. Amend, A.: Analyzing a terrorist's social media manifesto: The Pittsburgh synagogue shooter's posts on Gab. Southern Poverty Law Center, October 28. <https://www.splcenter.org/hatewatch/2018/10/28/analyzing-terrorists-social-media-manifesto-pittsburgh-synagogue-shooters-posts-gab> (2018). Accessed 13 Sep 2022
28. Institute of Medicine: *The Future of Public Health*. National Academies Press, Washington (1988)
29. Kass, N.E.: An ethics framework for public health. *Am. J. Public Health* **91**(11), 1776–1782 (2001). <https://doi.org/10.2105/ajph.91.11.1776>
30. Childress, J.F., Faden, R.R., Gaare, R.D., et al.: Public health ethics: mapping the terrain. *J. Law Med. Ethics* **30**, 170–178 (2002). <https://doi.org/10.1111/j.1748-720x.2002.tb00384.x>
31. Rothstein, M.A.: Rethinking the meaning of public health. *J. Law Med. Ethics* **30**, 144–149 (2002). <https://doi.org/10.1111/j.1748-720x.2002.tb00381.x>
32. Verweij, M., Dawson, A.: The meaning of 'public' in 'public health.' In: Dawson, A., Verweij, M. (eds.) *Ethics, Prevention, and Public Health*, pp. 13–29. Oxford University Press, Oxford (2007)
33. Detels, R., Tan, C.C.: The scope and concerns of public health. In: Detels, R., Gulliford, M., Karim, Q.A., Tan, C.C. (eds.) *Oxford Textbook of Global Public Health*, pp. 3–18. Oxford University Press, Oxford (2015)
34. Shultz, J.M., Zakrisson, T.L., Galea, S.: Hate and the health of populations. *Milbank Q.* **97**(1), 11–15 (2019). <https://doi.org/10.1111/1468-0009.12362>
35. Centers for Disease Control and Prevention.: About social determinants of health (SDOH). Last modified March 10. <https://www.cdc.gov/socialdeterminants/about.html> (2021). Accessed 20 Sep 2022
36. American Public Health Association.: Racism and health. <https://www.apha.org/Topics-and-Issues/Health-Equity/Racism-and-health> (2022). Accessed 22 July 2022
37. NYC Office for the Prevention of Hate Crimes. (2022). <https://www.nyc.gov/site/stophate/index.page>. Accessed 9 Nov 2022
38. Movement Advancement Project.: Policy spotlight: hate crime laws. Last Modified July 2021. <https://www.lgbtmap.org/2021-report-hate-crimes>. Accessed 14 Nov 2022
39. Gaffney, H., Farrington, D.P., Tfofi, M.M.: Examining the effectiveness of school-bullying intervention programs globally: a meta-analysis. *Int. J. Bullying Prev.* **1**, 14–31 (2019). <https://doi.org/10.1007/s42380-019-0007-4>
40. Butts, J.A., Roman, C.G., Bostwick, L., Porter, J.R.: Cure violence: a public health model to reduce gun violence. *Annu. Rev. Public Health* **36**, 39–53 (2015). <https://doi.org/10.1146/annurev-publhealth-031914-122509>
41. Williams, M.L., Burnap, P., Javed, A., Liu, H., Ozalp, S.: Hate in the machine: anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *Br. J. Criminol.* **60**, 93–117 (2020). <https://doi.org/10.1093/bjcz/azz049>
42. Schroepfer, M. Update on our progress on AI and hate speech detection. Meta, February 11. <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/> (2021). Accessed 8 July 2022
43. Twitter.: Hateful conduct policy. <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> (2022). Accessed 8 July 2022
44. Konikoff, D.: Gatekeepers of toxicity: Reconceptualizing twitter's abuse and hate speech policies. *Policy Internet* **13**, 502–521 (2021). <https://doi.org/10.1002/poi3.265>
45. Dias Oliva, T., Antonialli, D.M., Gomes, A.: Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online'. *Sex Cult.* **25**, 700–732 (2021). <https://doi.org/10.1007/s12119-020-09790-w>
46. ElSherief, M., Ziems, C., Muchlinskiy, D., Anupindi, V., Seybolty, J., De Choudhury, M., Yang, D.: Latent hatred: a benchmark for understanding implicit hate speech. *Empir. Methods Natl. Lang. Process.* (2021). <https://doi.org/10.18653/v1/2021.emnlp-main.29>
47. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. *Proc. Annu. Meet. Assoc. Comput. Linguist.* (2019). <https://doi.org/10.18653/v1/P19-1163>
48. Waseem, Z.: Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter. *Proc. First Worksh. NLP Comput. Soc. Sci.* (2016). <https://doi.org/10.18653/v1/W16-5618>
49. Binns, R., Veale, M., Van Kleek, M., Shadbolt, N.: Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: Ciampaglia, G., Mashhadi, A., Yasseri, T. (eds.) *Social Informatics*, pp. 405–415. Springer, Cham (2017)
50. Albadi, N., Kurdi, M., Mishra, S.: Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. *Proc. ACM Hum. Comput. Interact.* **3**, 1–25 (2019)
51. Davis, C. A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: BotOrNot: a system to evaluate social bots. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 273–274 (2016)
52. Uyheng, J., Carley, K.M.: Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *J. Comput. Soc. Sci.* **3**, 445–468 (2020). <https://doi.org/10.1007/s42001-020-00087-4>

53. Bailurkar, R., Raul, N.: Detecting bots to distinguish hate speech on social media. In: 12th International Conference on Computing Communication and Networking Technologies, pp. 1–5 (2021)
54. Uyheng, J., Bellutta, D., Carley, K.M.: Bots amplify and redirect hate speech in online discourse about racism during the COVID-19 pandemic. *Soc. Media Soc.* (2022). <https://doi.org/10.1177/20563051221104749>
55. Kovacs, G., Alonso, P., Saini, R.: Challenges of hate speech detection in social media. *SN Comput. Sci.* **2**(95), 2–12 (2021). <https://doi.org/10.1007/s42979-021-00457-3>
56. Cypris, N. F., Engelmann, S., Sasse, J., Grossklags, J., Baumert, A.: Intervening against online hate speech: A case for automated Counterspeech. In: IEAI Research Brief, pp. 1–8 (2022)
57. Mathew, B., Saha, P., Tharad, H., et al.: Thou shalt not hate: countering online hate speech. *Int. AAAI Conf. Web Soc. Med.* (2019). <https://doi.org/10.48550/arXiv.1808.04409>
58. Hangartner, D., Gennaro, G., Alasiria, S., et al.: Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proc. Natl. Acad. Sci. USA* **118**(50), 1–3 (2021). <https://doi.org/10.1073/pnas.2116310118>
59. Garland, J., Ghazi-Zahedi, K., Young, J.-G., Herbert-Dufresne, L., Galesic, M.: Countering hate on social media: large-scale classification of hate and counter speech. *Proc. Worksh. Online Abuse Harms* (2020). <https://doi.org/10.18653/v1/P17>
60. Garland, J., et al.: Impact and dynamics of hate and counter speech online. *EPJ Data Sci.* (2022). <https://doi.org/10.1140/epjds/s13688-021-00314-6>
61. de los Riscos, A.M., Haro, L.F.: ToxicBot: a conversational agent to fight online hate speech. In: Haro, L.F., Callejas, Z., Nakamura, S. (eds.) *Conversational Dialogue Systems for the Next Decade. Lecture Notes in Electrical Engineering*, pp. 15–29. Springer, Singapore (2021)
62. Arango, A., Perez, J., Poblete, B.: Hate speech detection is not as easy as you may think: a closer look at model validation. In: 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 45–54 (2019)
63. Savage, M.: Kanye West's Twitter and Instagram accounts locked over anti-Semitism. *BBC*. <https://www.bbc.com/news/entertainment-arts-63198991> (2022). Accessed 24 Oct 2022
64. Da Silva, C., Dasrath, D.: Rise in antisemitism is feared after banner saying 'Kanye is right' is hung over Los Angeles freeway. *NBS News*. <https://www.nbcnews.com/news/us-news/banner-kanye-right-los-angeles-freeway-antisemitic-group-rcna53653> (2022). Accessed 14 Nov 2022
65. Kim, J.: Elon Musk says Twitter restored Ye's account without his knowledge before acquisition. *NPR*. <https://www.npr.org/2022/10/28/1132222606/twitter-account-elon-musk-content-moderation-council> (2022). Accessed 14 Nov 2022
66. Gostin, L.O.: Public health: a renaissance. *J. Law Med. Ethics* **30**(2), 136–140 (2002)
67. Hosny, A., Parmar, C., Quackenbush, J., et al.: Artificial intelligence in radiology. *Nat Rev Cancer* **18**, 500–510 (2018). <https://doi.org/10.1038/s41568-018-0016-5>
68. Morgenstern, J.D., Rosella, L.C., Daley, M.J., Goel, V., Schunemann, H.J., Piggott, T.: “AI’s gonna have an impact on everything in society, so it has to have an impact on public health”: a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. *BMC Public Health* **21**(40), 1–14 (2021). <https://doi.org/10.1186/s12889-020-10030-x>
69. Bogoch, I.L., et al.: Anticipating the international spread of Zika virus from Brazil. *Lancet* **387**, 335–336 (2016). [https://doi.org/10.1016/S0140-6736\(16\)00080-5](https://doi.org/10.1016/S0140-6736(16)00080-5)
70. Resnik, D.B.: Responsibility for health: personal, social, and environmental. *J. Med. Ethics* **33**(8), 444–445 (2006). <https://doi.org/10.1136/jme.2006.017574>
71. Purnat, T.D., et al.: Infodemic signal detection during the COVID-19 pandemic: development of a methodology for identifying potential information voids in online conversations. *JMIR Infodemiology* **1**(1), 1–12 (2021). <https://doi.org/10.2196/30971>
72. Tangcharoensathien, V., et al.: Framework for managing the COVID-19 infodemic: methods and results of an online, crowd-sourced WHO technical consultation. *J Med Internet Res* **22**(6), 1–8 (2020). <https://doi.org/10.2196/19659>
73. Meltwater.: social listening. <https://www.meltwater.com/en/products/social-media-monitoring> (2023). Accessed 23 Feb 2023
74. Purnat, T.D., Wilson, H., Nguyen, T., Briand, S.: EARS: a WHO platform for AI-supported real-time online social listening of COVID-19 conversations. *Stud. Health Technol. Inf.* **281**, 1009–1010 (2021). <https://doi.org/10.3233/SHTI210330>
75. World Health Organization.: Early AI-supported response with social listening. <https://www.who-ears.com/> (2022). Accessed 5 August 2022
76. Citibeats.: The citibeats platform. <https://www.citibeats.com/platform> (2023). Accessed 24 Feb 2023
77. World Health Organization.: Methodology, data and definitions of the WHO early AI-supported response with social listening platform. <https://www.who-ears.com/#/methodology> (2023). Accessed 23 Feb 2023
78. Matamoros-Fernandez, A., Farkas, J.: Racism, hate speech, and social media: a systematic review and critique. *Telev. New Media* **22**(2), 205–224 (2021). <https://doi.org/10.1177/1527476420982230>
79. Johnson, A., Castro, D.: Fact checking the critiques of section 230: what are the real problems?. *Information Technology and Innovation Foundation*. <https://itif.org/publications/2021/02/22/fact-checking-critiques-section-230-what-are-real-problems/> (2021). Accessed 24 Feb 2023
80. Abroms, L.C.: Public health in the era of social media. *Am. J. Public Health* **109**(2), 130–131 (2019). <https://doi.org/10.2105/AJPH.2018.304947>
81. Stockmann, D.: Tech companies and the public interest: the role of the state in governing social media platforms. *Inf. Commun. Soc.* (2022). <https://doi.org/10.1080/1369118X.2022.2032796>
82. Hintz, A.: Restricting digital sites of dissent: commercial social media and free expression. *Crit. Discourse Stud.* **13**(3), 325–340 (2016). <https://doi.org/10.1080/17405904.2016.1141695>
83. Lauer, D.: Facebook’s ethical failures are not accidental; they are part of the business model. *AI Ethics* **1**, 395–403 (2021). <https://doi.org/10.1007/s43681-021-00068-x>
84. Morley, J., Cowls, J., Taddeo, M., Floridi, L.: Public health in the information age: recognizing the infosphere as a social determinant of health. *J. Med. Internet Res.* (2020). <https://doi.org/10.2196/19311>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.