AI-based removal of hate speech from digital social networks: chances and risks for freedom of expression

Frank Dietrich

Article - Version of Record

Suggested Citation:
Dietrich, F. (2024). AI-based removal of hate speech from digital social networks: chances and risks for freedom of expression. AI and Ethics, 5(3), 2943–2953. https://doi.org/10.1007/s43681-024-00610-7

AI and Ethics (2025) 5:2943–2953
https://doi.org/10.1007/s43681-024-00610-7

ORIGINAL RESEARCH

AI-based removal of hate speech from digital social networks: chances and risks for freedom of expression

Frank Dietrich[1]

Received: 2 May 2024 / Accepted: 3 November 2024 / Published online: 21 November 2024

Abstract
Given the enormous number of posts, major digital social networks, such as Facebook, must rely on artificial intelligence (AI) systems to regulate hate speech. This article explores the risks for free speech that the automated deletion of posts entails and discusses how AI systems can be subjected to human control. In a first step, the article examines the relevance of the individual right to freedom of expression for privately operated Internet platforms. It then highlights the specific risks that arise when AI systems are entrusted with the task of identifying and removing hate speech. The recently passed EU AI Act represents the most ambitious attempt to date to regulate high-risk AI applications. The article examines whether and, if so, to what extent the various forms of human oversight mentioned in the EU AI Act are feasible in the area of hate speech regulation. Three core theses are put forward: First, the deletion of hate speech by AI systems constitutes a high-risk application that requires an extension of the regulatory scope of the EU AI Act. Second, ex-post monitoring is the only feasible kind of human supervision but fails to guarantee full protection of the individual right to freedom of expression. Third, despite this shortcoming, the implementing of ex-post monitoring is necessary and legitimate to curb hate speech on digital social networks.

Keywords Artificial intelligence · Digital social networks · EU AI Act · Freedom of expression · Hate speech · Human oversight

1 Introduction

The first digital social networks that started to operate on the Internet in the 1990s largely refrained from imposing speech norms on their users. The ideal of an open and free cyberspace, where almost anything could be said, remained dominant in the years to follow. However, the major digital social networks were increasingly criticized for creating toxic environments especially for members of vulnerable groups. Due to mounting public pressure and the fear of being subjected to legal regulations, Facebook and other influential providers changed their policies in the mid-2010s. They committed themselves to banning hate speech on their platforms and tightened their community standards accordingly.[1] Thereby, they moved away from the American legal tradition of a widely unconstrained freedom of expression as it is enshrined in the First Amendment of the U.S. Constitution. Instead, they approached the European legal perspective which views the prohibition of demeaning and possibly intimidating speech as a prerequisite for the exercise of free speech by everyone.

The enormous amount of contributions posted every day by millions of users worldwide poses a great challenge for online content moderation. AI systems play an important role in the efforts of all major digital social networks to eliminate hate speech or to reduce its visibility. Currently, many different speech control tools based on machine learning are already in use. AI systems can, for instance, be used to flag potentially offensive posts in order to warn users against insults or slurs. Furthermore, search engines or recommendation systems can deprioritize problematic content

✉ Frank Dietrich
Frank.Dietrich@hhu.de

[1] Heinrich-Heine-University Düsseldorf, Institute of Philosophy, Universitätsstraße 1, 40225 Düsseldorf, Germany

[1] After Elon Musk, who portrays himself as a "free speech absolutist", purchased Twitter in October 2022 (and later renamed it X), the company's efforts to eliminate hate speech have declined significantly [1].

with the aim of containing its dissemination.[2] Digital social networks can also apply AI systems to detect presumably harmful content and to delegate the relevant posts to human review. Finally, algorithms can be the basis for blocking the upload of potential hate speech, for removing content, or even for the deletion of user accounts without any human involvement in the decision-making [3].

Currently, leading digital social networks, such as Facebook and Instagram, operate with hybrid systems in which artificial intelligence makes independent decision but also supports human reviewers. In a recent transparency report, they state: "To enforce our Community Standards, we employ a combination of human review and technology (…). Every day, we remove millions of violating pieces of content and accounts on Facebook and Instagram. In most of the cases, this happens automatically, with technology such as artificial intelligence working behind the scenes to detect and remove Community Standards violating content. In other cases, our technology selects content for human review" [4]. Moreover, Facebook apparently endeavors to further develop and refine intelligent software capable of proactively preventing hate speech from being posted. Therefore, the future might well belong to content moderation systems that fully dispense with human control and solely rely on AI solutions [5].

Obviously, a significant part of social communication today takes place within a small number of very influential digital social networks operating on a global scale. Typically, they are not much constrained by national legislation but enjoy extensive freedom to establish their own speech rules. Their operation of AI systems to eliminate hate speech has the potential to shape the public discourse and carries the risk of violating individual rights of expression. With regard to profit-oriented companies that have to compete in the market, it cannot be taken for granted that the protection of individual rights is their primary concern. Therefore, the question arises as to whether the use of AI systems in digital social networks needs to be legally regulated and, if so, what their control might look like.

In the following section I will dwell on the relevance of the right to free speech for digital social networks and their efforts to combat offensive posts. In the third section I will highlight the particular challenges and dangers of entrusting AI systems with the blocking or deletion of hate speech. Subsequently, in the fourth section I will discuss the European Union's ambitious legislative initiative to subject potentially harmful AI applications to human supervision. In the fifth section I will address the feasibility of human oversight over the automated removal of hate speech and scrutinize the legitimacy of—what I believe is—the only viable form of control. In the concluding section I will briefly summarize the main findings of my investigation.

In this article I will essentially put forward three theses. I will argue, first, that the deletion of allegedly offensive posts constitutes a high-risk application of AI systems requiring an extension of the areas covered by the EU AI Act. Second, I will maintain that the ex-post monitoring of an AI system's overall activity is the only feasible kind of human supervision. However, ex-post monitoring fails to provide comprehensive protection for the individual's right to freedom of expression. I will, third, defend the view that the implementation of this control tool is necessary and legitimate despite its shortcomings.

2 Freedom of expression in digital social networks

The right to freedom of expression encompasses an active and a passive element: it includes the right to share one's own opinion with other people and the right to learn about other people's views. Both the active and the passive components of the right to freedom of expression protect fundamental individual interests. As social beings, humans typically have a strong desire to communicate with their fellows and to disclose their feelings and thoughts to them. In general, they also benefit greatly from participating in the knowledge of others and considering their—often divergent—opinions [6]. Moreover, the right to freedom of expression fulfills a vital function in democratic societies, by enabling citizens to engage in public deliberation. Informed democratic decisions, which are responsive to the will of the electorate, can only be made if all members of the political community are able to voice their values and interests [7].

The enormous expansion of communicative possibilities offered by the Internet and digital social networks in particular has had a major impact on the freedom of expression. In the analogue age only a few people, especially journalists and politicians, were able to reach a wide audience with their messages. Before the advent of the Internet most people could only share their views in their private spheres with a narrowly defined group of people. Digital social networks have basically made it possible for everybody to connect with a large number of people around the world. Whereas previously most people were limited to the role of passive readers or listeners, they have now become "potential authors". As Jürgen Habermas recently noted, digital social networks have thus contributed significantly to a new structural transformation of the public sphere [8].

A serious problem with Internet communication is, of course, that the quality control which characterizes

[2] Restricting the visibility of potentially hateful content is at the heart of the "freedom of speech, not reach" policy that X has recently introduced [2].

professional journalistic contributions is lacking. In addition, the emergence of echo chambers and the resulting fragmentation of public deliberation pose a threat to social cohesion and the acceptance of democratic institutions [8]. Regardless of these challenges, however, it is important to keep in mind the freedom-promoting achievements of digital social networks. They have, as Thiago Dias Oliva put it, taken "freedom of expression and access to information to a whole new level of enjoyment" [9]. The argument for the application of AI systems which I will advance in the fifth section is based primarily on the enhanced opportunities digital social networks offer for free speech.

The individual fundamental rights enshrined in the constitutions of most democratic countries establish, in the first place, claims against the state. Thus, the right to freedom of expression primarily protects citizens from state prohibitions but does not—or only under specific conditions—extend to the private sphere. For instance, if Susan excludes Oliver from her garden party because she strongly dislikes his political views, she does not violate his freedom of expression. Oliver has neither any right to be invited to Susan's party, nor to express his political opinion in her private garden. By contrast, a city council is not authorized to ban persons who hold political or religious views that others find disturbing from markets or other public places. Basically, every citizen has the right to enter public spaces and to express their personal beliefs to everybody who is willing to listen.

Since the public-private distinction is crucial for the assessment of possible violations of rights, it is important to clarify in which sphere digital social networks operate. The removal of alleged hate speech by or with the help of artificial intelligence would only affect the right to freedom of expression if digital social networks could be considered part of the public sphere. Evidently, the most influential digital social networks, such as Facebook, YouTube, WhatsApp or Instagram, are run by private companies. Therefore, it may be tempting to think that their efforts to eliminate offensive or discriminatory posts are not constrained by the individual right to freedom of expression. However, strong arguments can be advanced as to why at least the major digital social networks have become a constitutive element of public communication in democratic societies.

The stated goal of Facebook and other digital social networks is not to connect a specific group of people, such as hobby gardeners or chess players. Instead, they have been aiming at providing open platforms for everybody and have thereby created—with great success—a virtual public sphere.[3] In recent decades, social communication has increasingly shifted from traditional public places, such as markets or town halls, to the Internet. Today, digital social networks have become the most important place for many people to express their opinions, gather information and exchange ideas with others. Moreover, to a wide extent, states have left the regulation of Internet communication to the providers themselves. Facebook and other companies in a sense act as law-givers whose terms of service largely determine the virtual speech norms. Therefore, the most influential digital social networks should be considered a "quasi-public forum" to which the individual right to freedom of expression applies [12, 13].[4]

Of course, the question as to what extent, if at all, the regulation of hate speech is consistent with the freedom of expression sparks much controversy. According to Jeremy Waldron, a ban on derogatory posts can be necessary to protect public order which he understands in a narrow and a broad sense [15]. In a narrow sense, public order is characterized by safety from physical attacks and can be disturbed if the use of, for instance, fighting words leads to violence. In a broad sense, public order also requires that every citizen can be sure of being recognized as a full member of society with the same rights as everyone else. In particular, group defamation that questions the equal status of people who share certain characteristics, such as skin color, undermines the certainty of being respected and protected. As Waldron [15] puts it, hate speech laws contribute to maintaining "public order, not just by preempting violence, but by upholding against attack a shared sense of the basic elements of each person's status, dignity, and reputation as a citizen or member of society in good standing (…)".

According to Waldron, the expression of hate speech has the potential to shape the real or virtual social environment of the persons concerned. As a consequence of knowing about the extreme animosity of other citizens, the members of the targeted group often feel threatened and intimidated. In addition, the permanent exposure to humiliating statements may significantly impair the self-assurance and self-esteem of those who are denied equal status. Although Waldron identifies important harms caused by hate speech, its potentially silencing effects deserve even more attention. If hate speech is not adequately curbed, the members of discriminated groups may be afraid of offensive responses and may prefer to remain "invisible". A sense of discomfort or

[3] Please note that the argument in favor of hate speech regulation I will advance in this section does not apply to messaging groups whose membership is typically limited to some hundred or thousand users by the leading providers [10]. Admittedly, however, the distinction from the public sphere can be difficult when messaging groups consist of up to 200.000 members, such as allowed by Telegram [11].

[4] The private character of digital social networks has also been successfully challenged with regard to politicians, most notably Donald Trump, who disseminate political messages via their accounts. If they use X or other networks to communicate with their constituents, they thereby create a "quasi-public forum" and are not allowed to block users who write critical comments [14].

fear may prevent them from speaking out in public, which also deprives possible audiences of the opportunity to learn about their views. In addition, hate speech can have the more subtle effect of making the statements of persons belonging to stigmatized groups count for nothing (or at least not for what they wanted to express). If certain groups are perceived in a very negative light, e.g. as being untrustworthy or vicious, it can be extremely difficult for their members to be understood in the way they intend.[5]

The considerations so far have shown that the expression of hate speech by some people may prevent other people from expressing their views. Since every citizen has the same right to freedom of expression, there are not only external reasons, such as protection from physical attacks that can be triggered by defamation or slander, for regulating hate speech. States also have an internal reason, arising from their obligation to ensure the right to freedom of expression on an equal basis to counteract the potentially silencing effects of disrespectful speech [16, 18]. In addition, the passive element of the freedom of expression, i.e. the right to learn of the views of others, also requires that every citizen be able to speak out without fear. A more or less extensive ban on hate speech may be the only way to maintain a social environment where everybody feels safe enough to contribute to public debates. This seems to be especially true for digital social platforms where provocative statements tend to attract most attention and are easily shared with countless other users.

In the following I will not address in any detail the complicated question of which criteria must be met in order to classify a statement as hate speech. I merely presume that at least in some cases a prohibition of threatening, insulting, or demeaning language can plausibly be substantiated. The focus of this paper is on the specific risks that arise when hate speech on digital social networks is regulated by or with the help of AI. To discuss the particular challenges posed by AI systems, it is not necessary to clarify the exact scope of permissible restrictions. The problems that will be examined in the next section arise regardless of the extent to which free speech is interfered with. The inquiry would only lose its point if any way of hate speech elimination had to be considered illegitimate.

Since I make no specific assumptions about the appropriate scope of regulation, my reflections are relevant to a wide range of theoretical perspectives and legal systems. Even theorists, like Matthew Kramer, who are highly critical of speech limitations typically allow for a ban on certain classes of expression, such as targeted harassment [12]. Only a few authors, such as Eric Heinze, argue that in "longstanding, stable and prosperous democracies" no curtailment of citizens' right to say whatever they want is warranted [19]. Furthermore, to my knowledge, no democratic state leaves the public communication of its citizens completely uncensored. Even the First Amendment of the U.S. constitution, which provides a comparatively high level of protection for individual free speech, allows for some exceptions, e.g. regarding obscenity or fighting words [20].

3 The risks of using artificial intelligence to protect against hate speech

Although AI-driven content moderation systems have improved tremendously in recent years, they are still far from being perfect. Especially in the area of hate speech detection and removal, machine learning approaches face significant challenges. Most AI systems currently in use rely on training data that already contains verified examples of hate speech.[6] For this purpose, human reviewers—trained experts or often crowd workers—have to decide whether posts are toxic or innocuous. Computer analysis then allows to identify potentially complex statistical relationships between linguistic features and the statements that are classified as hate speech. Based on the recognized patterns, the machine learning system can predict whether a still unknown text includes offensive language.

A major problem for the detection of hate speech machine learning systems face is the variability and context-dependency of linguistic meaning [22, 23]. For example, terms that typically indicate an offensive or discriminatory statement may be used in a completely different sense in certain subcultures. Minority groups sometimes use expressions that would normally be considered insults as greetings or in other everyday usage to signal membership with the group. Furthermore, in some cases they deliberately reclaim discriminatory terms that have traditionally been directed against the group by giving them a positive political meaning. On the other hand, also those who spread hate on the Internet may use specific codes or introduce new labels to avoid the removal of their posts. Apart from the different meanings language can have in group-specific discourses,

[5] The latter aspect of the "silencing effect" has been explored in depth by feminist authors, such as Rae Langton and Ishani Maitra, in the context of the debate on pornography [16, 17]. In their view, the constant portrayal of women as willing objects of sexual intercourse makes it difficult, if not impossible, for their "no" to be understood as "no".

[6] In this paper I focus exclusively on AI-supported prediction of hate speech and leave moderation systems that aim at matching content with the help of "hashes" or blacklists out of consideration. Both matching technologies have significant disadvantages for detecting offensive language: while "hashes" are vulnerable to minor changes of content, it is difficult to keep blacklists up to date [21].

ironic, sarcastic, or parodic statements also pose major challenges to AI systems [21].

AI systems designed to detect and delete hate speech can err in a variety of ways. On the one hand, they may take false negative decisions by classifying offensive or discriminatory posts as innocuous. At a first glance, the failure to detect hate speech may seem to pose no threat for the freedom of expression, as it does not result in the deletion of posts. However, as outlined above, persons who are potentially exposed to hateful comments may be discouraged to express their views. Especially members of vulnerable groups may refrain from contributing to the public debate if they anticipate insulting or demeaning responses. Similarly, persons who hold opinions that are likely to provoke hostility on digital social platforms may refrain from posting. In addition, by silencing members of minority groups through hate speech, potential listeners are also prevented from engaging with their views. In sum, false negative judgments counteract the creation of a respectful social environment in which everyone can speak out safely. Therefore, they tend to undermine the goal of guaranteeing every citizen the right to freedom of expression on an equal basis.

A further problem may arise when the social groups that participate in public discourse are affected differently by false negative decisions. As an illustration, consider two hostile groups who frequently attack members of the other community with insulting and derogatory comments. Imagine further that both groups use different speech-codes that are not represented to the same extent in the training data. If, therefore, the training data contains only or predominantly examples of hate speech from one of the groups, the AI system will be much better able to detect violations of speech norms by members of this group. As a result, the false negative decisions that prevent the automatic detection system from removing posts will have a biased distribution. While members of one group are frequently confronted with the deletion of their posts, members of the other group are able to spread hate speech largely unhindered [21, 24]. The intimidation and possible silencing of the disadvantaged group contradicts the obligation to equally grant every citizen the right to freedom of expression. Moreover, the less frequent sanctioning of the rival group may create the impression of unfairness and undermine any still existing willingness to comply with laws against hate speech.

On the other hand, AI systems may take false positive decisions by categorizing inoffensive posts that do not bear any risk of silencing other citizens as hate speech. To the extent that the assessment of a statement as being toxic results in its automatic removal (or deprioritization), the freedom of expression is directly affected. The people concerned are prevented from sharing their views with others and are therefore effectively censored. The deletion of an unobjectionable post is first and foremost a violation of the speaker's right to freedom of expression. However, due to the passive aspect of the freedom of expression, it also constitutes an infringement of the rights of the potential addressees who are entitled to take note of the statement. Since hate speech elimination by or with the help of AI systems poses significant threats for a core individual freedom, it should not be left unregulated. In the following section I will discuss the EU's legislative efforts to control high-risk AI applications, which is currently the most ambitious initiative.

4 The EU legal framework

The EU's regulatory approach aims to promote the development of trustworthy AI to ensure a high level of acceptance among potential users. It thus offers an important alternative, inter alia to U.S. and Chinese policies, which largely forego binding legal requirements.[7] The EU seeks to achieve the trustworthiness of future AI applications through a "human-centric" approach that focuses on human interests and values.[8] The goal of retaining control over AI is already visible in the General Data Protection Regulation (GDPR) the EU adopted in May 2018. Article 22(1) of the GDPR grants everyone the "right not to be subject to a decision based solely on automated processing, including profiling, which produces legal affects concerning him or her or similarly significantly affects him or her." The wording of Article 22(1) leaves open in which ways human actors must be involved in decisions made by or with the help of AI systems. Furthermore, according to Article 22(2), the above-stated provision does not apply if the person whose data is processed has explicitly consented to the AI application. Nevertheless, the GDPR already contains in nuce a right to human participation in decision-making that may have serious consequences for the data subject [27].

The idea of trustworthy AI was further developed in the Ethics Guidelines presented by a high-level expert group of the EU in April 2019 [28]. The document lists seven key requirements AI systems are supposed to meet in order to be perceived as being trustworthy, the first of which is the

[7] The U.S. President's Executive Order 14,110 on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" from 30th October 2023 may, however, signal a realignment of U.S. policies [25].

[8] In a paper on the creation of trust in human-centric AI the European Commission states: „Ethical AI is a win-win proposition. Guaranteeing the respect for fundamental values and rights is not only essential in itself, it also facilitates acceptance by the public and increases the competitive advantage of European AI companies by establishing a brand of human-centric, trustworthy AI known for ethical and secure products" [26].

requirement for "human agency and oversight".[9] Under this headline the Ethics Guidelines distinguish three governance mechanisms—human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC)—through which human control over AI applications can be exercised: "HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation."

Human involvement in decision-making is also a fundamental concern of the EU AI Act proposed by the European Commission in April 2021. After the European Council and the European Parliament had presented their positions in December 2022 and June 2023 respectively, interinstitutional negotiations began in December 2023. The compromise reached in the trilogue process was formally adopted by the European Parliament in March 2024 and the European Council in May 2024. The regulation entered into force in August 2024, its provisions coming into operation gradually over the next six to thirty-six months. The EU AI Act takes a risk-based approach, prescribing legal interventions according to the risk-level of the respective application. Specifically, a distinction is made between AI systems that impose unacceptable risks, which are completely prohibited, high risks, limited risks and minimal risks. While the operation of systems in the latter two categories entails no or only transparency obligations, high-risk systems are subject to strict regulation.

Instead of providing a definition of the term "high risk", Art. 6(2) of the EU AI Act refers to Annex III which lists eight areas where artificial intelligence systems may pose serious threats [29]. Art. 7(1) of the EU AI Act authorizes the European Commission to supplement Annex III with applications where both of the following conditions are met: "(a) the AI systems are intended to be used in any of the areas listed in points 1 to 8 of Annex III; (b) the AI systems pose a risk of harm to health and safety, or an adverse impact on fundamental rights, and that risk is equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in Annex III" [29]. Thus, Art. 7(1) in combination with Art. 14(2) of the EU AI Act indicates that possible threats to health, safety or fundamental rights are the relevant consideration for classifying AI systems as being high risk.[10]

According to Art. 14(1) of the EU AI Act, "high-risk AI systems shall be designed and developed in such a way (…) that they can be effectively overseen by natural persons during the period in which the AI system is in use" [29]. Consequently, it is the responsibility of the provider to enable human oversight before placing a high-risk AI system on the market. Art. 14 (4) of the EU AI Act enumerates various options an AI system should grant to those who are assigned the task of supervising it. For instance, those concerned should "be able to duly monitor its operation, also in view of detecting and addressing anomalies, disfunctions and unexpected performance." They should be enabled "to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse [its] output." Moreover, they should be enabled "to intervene on the operation of the high-risk AI system, or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state."

Presumably, Art. 14(4) of the EU AI Act merely lists alternative ways in which human oversight of high-risk AI systems can be exercised. The introductory sentence of Art. 14(4) explicitly states that those responsible for oversight need only have the above-mentioned options "as appropriate and proportionate to the circumstances." This wording suggests that the requirements AI systems must meet are context-dependent and likely to vary considerably between different applications. Therefore, for each specific high-risk AI system it still needs to be clarified how it must enable the operator to exercise some kind of human supervision [30]. According to Art. 43 of the EU AI Act, providers of high-risk AI systems have to perform a conformity assessment that demonstrates full compliance with the legal requirements before placing their product on the market. In addition, Art. 61 and Art. 21 of the EU AI Act, respectively, oblige the providers of high-risk AI systems to establish post-market monitoring plans and to take corrective action if they detect malfunctions.

In the previous sections I have argued, first, that the protection afforded by the individual right to freedom of expression also applies to large digital social networks, even though they are run by private companies. Second, I maintained that due to the complexity and context-dependency of language, AI systems (at their present stage of development) bear a relatively high risk of misclassifying posts. False positive decisions, i.e. the deletion or deprioritizing

[9] In addition, the Ethics Guidelines refer to "technical robustness and safety"; "privacy and data governance"; "transparency"; "diversity, non-discrimination and fairness", "societal and environmental well-being", and "accountability" [28].

[10] Art. 14(2) of the EU AI Act states: "Human oversight shall aim at preventing or minimizing the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse (…)" [29].

of inoffensive posts, are tantamount to censorship of the users concerned. False negative decisions, i.e. the maintaining of discriminating or derogatory posts, contribute to the creation of a toxic atmosphere in which members of vulnerable groups in particular are effectively silenced. Therefore, the use of AI systems by digital social networks to eliminate hate speech poses a serious threat to the fundamental right to freedom of expression guaranteed by the EU and its member states [31].[11]

If my reasoning has been correct, AI systems which are responsible for detecting and removing hate speech should be considered high-risk applications requiring human oversight. However, the eight areas of high-risk systems listed in Annex III of the EU AI Act do not mention hate speech moderation within digital social networks [29]. A proposal by the EU Parliament to include recommender systems used by social media platforms in Annex III was not agreed with in the trilogue process [32]. Given that automated hate speech moderation constitutes a high-risk AI application, the question arises in which way human control should be exercised. As mentioned above, the EU AI Act identifies various ways of human oversight but remains rather vague about the concrete requirements specific AI systems should meet [33]. In the next section, I will discuss the feasibility and legitimacy of three options for subjecting automated hate speech moderation on digital social networks to human control. I will argue that the immense amount of posts that are uploaded to digital social networks on a daily basis only allows for a retrospective overall assessment of the system's functioning.

5 The implementation of human oversight

There are at least three approaches to implement human oversight over automated hate speech moderation that deserve discussion. A first option one might consider would be to subject every decision by an AI system to block or remove a post to human control. Obviously, such a human-in-the-loop approach would face the problem of having to deal with an enormous amount of automated eliminations of hate speech. Considering that the major digital social networks delete millions of posts every day, a human case-by-case revision appears to be infeasible. Trying to approach complete control by checking as many decisions by the AI system as possible does not seem promising either. Presumably, human reviewers would have to check a very large number of posts the AI system has classified as hate speech in a very short time. However, human actors typically perform very poorly at routine tasks, as they quickly suffer from fatigue and lack of concentration. Experience has shown that under great time pressure they tend to simply "rubber stamp" the AI's decisions without exercising any meaningful control [34, 35].

Given the enormous number of potentially toxic posts deleted by the major digital social networks a second, more realistic option would be to subject only a selection of cases to human review. Such a spot check may be seen as an instantiation of the human-on-the-loop approach which requires monitoring the operation of an AI system instead of single case control. When deciding which posts are to be delegated to human reviewers, a distinction could be made between standard cases which are relatively clear-cut and hard cases which are difficult to evaluate. If AI systems were only to decide on standard cases, they would exclusively perform those tasks in which they are—as things stand today—superior to human actors. As a result, the routine work of human reviewers would be greatly reduced, leaving them more time to assess problematic cases with due care.

The proposed distinction between standard cases and hard cases promises to combine machine efficiency with human attentiveness and judgment. However, the benefits of this approach can only be realized if both types of cases are properly distinguished and hard cases are reliably identified. The key question therefore is who is to decide about which cases are to be classified as being problematic and are to be delegated to human reviewers. Since every individual post needs to be sorted in the categories "hate speech", "no hate speech" or "hard case", it seems obvious that the AI system will have to do the work. This means that a crucial step in the procedure, namely the selection of posts for human review, would lack human oversight [3]. AI systems should not be expected to do a better job at recognizing hard cases than at evaluating them, as the context-dependency of language causes difficulties for both tasks. If, for example, certain key words qualify posts as standard cases of hate speech, their use in ironic statements or critical references to content may easily lead to misjudgments. Consequently, there is a serious risk that these posts will be automatically deleted without ever being evaluated by any human reviewer.

Another way to reduce the number of suspicious posts subject to human review would be to introduce a complaint procedure for users of digital social networks. In order to enable users to request a human review, they must be informed of the deletion of their post and of the possibility to lodge a complaint. In contrast to the previously discussed approach, the selection of cases for human review would not be made by an AI system but by the individuals whose posts

[11] Art. 11 (1) of the Charta of Fundamental Rights of the European Union states: "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and to impart information and ideas without interference by public authority and regardless of frontiers." See also the European Convention on Human Rights (ECHR), Art. 10 (1).

have been censored. An example of such a complaint procedure can be found in the German Network Enforcement Act (NetzDG) that came into force on October 1, 2017 [36]. The NetzDG obliges providers of digital social networks to delete evidently illegal content, e.g. posts that constitute an insult, defamation or libel under the German Criminal Code, within 24 h. Content that is unlawful, though not in an obvious way, must be removed within a period of seven days which can be extended under certain circumstances, as defined in § 3(3) NetzDG. Under the NetzDG, digital social networks are only obliged to respond appropriately to complaints; they cannot be penalized for hateful content to which no objections have been raised.[12]

Regarding appeal procedures, the Oversight Board (OB) which Facebook, now Meta, established in May 2020 as a kind of "supreme court" also deserves attention. The OB has an independent financial structure and an independent selection process for its up to forty members. According to Art. 2(1) of the Oversight Board Charter, users can submit a request for review if they disagree with the company's decision to delete or uphold controversial content and have exhausted appeals [37]. In addition, Meta itself can forward critical cases for review to the OB. The OB selects from thousands of submissions those cases it considers to "have the greatest potential to guide future decisions and policy" [37]. The decisions the OB is supposed to take with regard to Meta's community standards and human rights norms on freedom of expression are binding.[13] Furthermore, the OB may, on its own initiative or at Meta's request, issue non-binding policy advisory opinions to which the company must respond publicly.[14]

The complaint procedure provided by the NetzDG relies on human actors, viz. the objectors, being entrusted with the selection of the posts whose deletion has to be verified. However, granting every individual whose post has been classified as hate speech by the AI system a right to human revision raises its own difficulties. A complaint procedure can only work on the assumption that most people affected by the deletion of a post will not exercise their right to complain. The more people demand a human revision, the more difficult—and at a certain point impossible—it becomes to meet their requests. In an extreme scenario, in which everyone affected exercises their right to complain, the procedure would revert to a case-by-case review with all the problems mentioned above. The appeal procedure established by Meta leaves the selection of cases to a small group of OB's members who are only able to review a tiny fraction of the decisions considered questionable by users. When the OB began its work in October 2020, it received about 20.000 cases of which it accepted six for review in the first three months; as to April 2024 not more than 115 cases have been reassessed [40]. Although the OB promises to select cases that are systemically relevant, it clearly lacks the capacity to comprehensively handle the large number of complaints.

Another concern with providing a complaint or appeal procedure is that users may be very differently inclined to resort to it. Due to a lack of knowledge of how to file a complaint or a general distrust in the fairness of the process, some people may regularly refrain from requesting a human review. In particular, members of groups that are economically disadvantaged or have suffered from discrimination in the past may be less likely to exercise their right to complain. Hence, there is a risk that the protection of freedom of expression the availability of a review process is intended to guarantee will prove to be ineffective for some groups. Notwithstanding the aforementioned problems, complaint or appeal procedures can play—as I am going to argue shortly—some role in a wider auditing of AI systems. On their own, however, they seem to be insufficient to ensure adequate human oversight of the automated removal of hate speech in digital social networks.[15] They fall short of providing a comprehensive overview of existing problems, as they are only able to consider a small and potentially biased sample of cases. Moreover, complaint or appeal procedures only lead to a correction of single cases, but not necessarily to a re-training of the algorithm to avoid future errors.

A third and more promising method to achieve effective control is a human-in-command approach that aims at overseeing the overall activity of the AI system. This means that the operation of the entire AI system would have to be reviewed at regular intervals by an independent team of human experts. In contrast to Meta's initiative, which is based on a voluntary commitment, the requirement to establish some kind of oversight board should be enshrined in law. As indicated above, a complaint or appeal procedure could be an important element of the periodic monitoring of the AI system. Although obvious mistakes in the elimination of posts could be corrected immediately, the purpose of the complaint procedure would not be to grant every user a right to human review. The complaint or appeal procedure

[12] The NetzDG does not explicitly state that decisions on complaints must be made by human reviewers. However, the requirement mentioned in § 4 NetzDG, to provide regular training and support to the personnel entrusted with the complaint procedure, indicates that the legislator assumed human control.

[13] It should be noted, however, that the binding nature of the OB's decisions is based on a voluntary commitment by Meta, which can be revoked at any time [38]. The normative basis of the decisions is set out in Art. 2(2), and their obligatory effect for Meta in Art. 4 of the Oversight Board Charter [37].

[14] See Art. 1(4) of the Oversight Board Charter in connection with Art. 2(2.3.2) of the Oversight Board Bylaws [37, 39].

[15] For a proposal to combine an individual rights regime with some form of systemic oversight over AI-based decision-making, see Margot E. Kaminski [41].

would rather help to detect common errors and systematic biases and thus provide clues for the improvement of the AI system. In addition, a sufficiently large random sample of blocked or deleted posts would also have to be collected and evaluated by human reviewers. As the use of the complaint procedure may be highly distorted, a more representative selection of eliminated posts needs to be included in the monitoring of the AI system.

The aim of the monitoring process must be to gain a reliable impression of the overall decision-making of the AI system. In particular, it should serve for identifying typical errors in the classification and subsequent deletion of posts as hate speech. As a consequence of the monitoring, the digital social networks would be required to take corrective action on a systemic level by modifying the algorithm. The AI system could be improved, for example, by re-training it with specific data sets containing posts from minorities or subcultures that are often misjudged. Moreover, reinforcement learning from human feedback (RLHF) could be used to reduce the error rate in correctly identifying hate speech [42, 43].[16] The proposal outlined here is in line with the post-market monitoring process the EU AI Act requires from the providers of AI applications [29]. Although a human-in-command approach appears to be the only viable solution for implementing human control, it raises justificatory issues that need to be addressed.

The monitoring process faces the objection that it falls short of providing full protection of the right to freedom of expression as laid down in the major European human rights documents. Since oversight is exercised ex-post, it does not prevent current rights violations but only requires corrective action when malfunctions of the AI system have been identified. Furthermore, users whose posts have been blocked or eliminated are not granted any individual right to a human review and, if necessary, restoration of their content. The monitoring process merely obliges digital social networks to improve the effectiveness of the AI system in order to reduce the total number of unwarranted deletions in the future. Therefore, the question arises as to whether the implementation of a human-in-command approach can be justified despite its shortcomings. In the remainder of this section I will advance three interrelated arguments for the legitimacy of the monitoring process outlined above.

First of all, it should be emphasized that there is no better option to address the problem of hate speech on digital social networks. Considering the enormous number of contributions that are uploaded every day, it is impossible to guarantee full protection for every single post. Only AI-driven solutions which always go with a certain margin of error enable providers to moderate hate speech on a sufficiently large scale. As Evelyn Douek put it: "(…) Probabilistic enforcement is the *only* possibility between two extremes of severely limiting speech or letting all the posts flow" [45]. If every single violation of freedom of expression were to be avoided, the operation of digital social networks would have to be legally banned. However, such an extreme measure would be highly counterproductive for the advancement of free speech. "The internet has enabled more broadcasting of expression and amplification of speech, including by those historically marginalized and excluded, than any time in history. Platforms are powerful venues for free expression and a world without them would be a loss for free speech" [45].

On the other hand, digital social networks can only play a positive role for freedom of expression if the proliferation of hate speech is adequately curbed. The other extreme referred to by Douek, namely "letting all the posts flow", also has to be avoided in order to create an environment within which citizens feel safe to participate in the public discourse. As argued in the second section, exposure to derogating or discriminating posts can intimidate those targeted, especially members of vulnerable groups. The freedom of expression in both its active and passive dimensions can only be guaranteed to every citizen equally, if silencing effects are sufficiently contained. Hate speech by some individuals must not prevent others from stating their opinion or receiving information about a wide range of views. Consequently, there is no alternative to striving for continuous improvement of AI systems, knowing complete legal protection cannot be achieved in the foreseeable future.

Second, the infringement of the right to free speech that must be tolerated within a human-in-command approach is less severe than classical cases of censorship. Of course, every unwarranted elimination of a post constitutes a violation of the concerned user's rights that should not be trivialized. However, users will in general only be occasionally and arbitrarily affected by erroneous decisions by the AI system. Typically, they will still have ample opportunity to exercise their freedom of expression and to communicate their views to other people. By contrast, classical forms of censorship either prohibit certain persons from speaking out in public or penalize the public utterance of certain opinions, e.g. criticism of religion. Certainly, the more the AI system exhibits a systematic bias against specific groups or opinions, the closer it comes to the classical kind of censorship. However, ex-post monitoring aims precisely to solve this problem by re-training the AI system and improving its future functioning.

[16] For a critical evaluation of RLHF, emphasizing the need to represent a wide range of different human values in the feedback process, see McIntosh et al. [44]

Finally, the human-in-command solution I propose in this section offers an important advantage over the classical human rights view that deserves mentioning. The classical human rights view leaves only a choice between two options, namely deleting or upholding posts, depending on whether they are considered hate speech or tolerable expression. As I have argued in the second section, hate speech moderation should aim at protecting public order in a broad sense, by providing every citizen with the security of being recognized as being equal. For this reason, it is important to create a safe environment within which members of vulnerable groups in particular are not permanently exposed to discriminatory or derogatory comments. However, there are a variety of options for preventing hate speech from dominating the public discourse and potentially having a silencing effect. Besides the removal of hate speech, less severe measures to reduce its visibility, such as deprioritization or downgrading, are also available. Thus, my approach allows for the consideration of a wider range of technological options and promises to enable a more balanced and flexible form of hate speech moderation [45].

6 Conclusion

Given the enormous number of posts that are uploaded every day, there is no alternative to using AI systems to regulate hate speech. As the deletion of posts is an infringement of the fundamental right to freedom of expression and as AI-systems, at their current stage of development, are still prone to error, their operation requires human supervision. Therefore, the recently adopted EU AI Act, which fails to include the automated regulation of hate speech on digital social networks into the high-risk applications, should be amended. A critical scrutiny of various alternatives referred to in the EU AI Act (and other European legal documents) has shown that ex-post monitoring of the AI system's overall activity is the only viable option for exercising control. However, ex-post monitoring falls short of ensuring the protection of individual rights as it is understood in the classical human rights discourse. The justification for the application of AI systems to regulate hate speech must rely on a consequentialist argument based on the increased opportunities for free speech offered by digital social networks.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Frenkel, S., Conger, K.: Hate speech's rise on Twitter is unprecedented, researchers find. New York Times. (2022). https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html# Accessed 2 October 2024
2. Twitter, D.S.A.: transparency report (August 28th to October 20th 2023). (2024). https://transparency.twitter.com/dsa-transparency-report.html Accessed 2 October 2024
3. Wu, T.: Will artificial intelligence eat the law? The rise of hybrid social-ordering systems. Colum L Rev. **119**, 2001–2028 (2019). https://columbialawreview.org/content/will-artificial-intelligence-eat-the-law-the-rise-of-hybrid-social-ordering-systems/
4. Facebook NetzDG transparency report: (2022). https://about.fb.com/de/wp-content/uploads/sites/10/2022/07/Facebook-NetzDG-Transparency-Report-July-2022.pdf Accessed 2 October 2024
5. Transcript of Mark Zuckerberg's Senate hearing: (2018). https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/ Accessed 2 October 2024
6. Schauer, F.: Free Speech: A Philosophical Enquiry. Cambridge University Press, Cambridge (1984)
7. Meiklejohn, A.: Free Speech and its Relation to self-government. Harper & Brothers, New York (1948)
8. Habermas, J.: Ein Neuer Strukturwandel Der Öffentlichkeit Und die Deliberative Politik. Suhrkamp, Berlin (2022)
9. Oliva, T.D.: Content moderation technologies: Applying human right standards to protect freedom of expression. Hum. Rights Law Rev. **20**, 607–640 (2020). https://doi.org/10.1007/s12119-020-09790-w
10. WhatsApp: How to create a community. (2024). https://faq.whatsapp.com/438859978317289/?helpref=faq_content%26cms_platform=web Accessed 2 October 2024
11. Telegram: Groups and channels. (2024). https://telegram.org/faq?setln=en#groups-and-channels Accessed 2 October 2024
12. Kramer, M.H.: Freedom of Expression as self-restraint. Oxford University Press, Oxford (2021)
13. Shefa, M.C.: First amendment 2.0: revisiting *Marsh* and the quasi-public forum in the age of social media. U Haw L Rev. **41**, 159–188 (2018)
14. Morris, P.L., Sarapin, S.H.: You can't block me: When social media spaces are public forums. First Amendment Stud. **54**, 52–70 (2020). https://doi.org/10.1080/21689725.2020.1742760
15. Waldron, J.: The harm in hate Speech. Harvard University Press, Cambridge (2012)
16. Langton, R.: Speech acts and unspeakable acts. Phil Pub Aff. **22**, 293–330 (1993)
17. Maitra, I.: Silencing speech. Can. J. Philos. **39**, 309–338 (2009)
18. Howard, J.W.: Freedom of speech. In: Zalta, E. N., Nodelman, U. (eds.) Staenford Encyclopedia of Philosophy (Spring 2024

Edition). (2024). https://plato.stanford.edu/entries/freedom-speech/ Accessed 2 October 2024
19. Heinze, E.: Hate Speech and Democratic Citizenship. Oxford University Press, Oxford (2016)
20. Ruane, K.A.: Freedom of speech and press: Exceptions to the first amendment. Library of Congress. (2014). https://digital.library.unt.edu/ark:/67531/metadc462149/#partners Accessed 2 October 2024
21. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. **7**, 1–15 (2020). https://doi.org/10.1177/2053951719897945
22. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. **51**, Article 85 ((2018). https://doi.org/10.1145/3232676
23. Carstens, J.A.: D. Friess AI within online discussions: Rational, civil, privileged? Ethical considerations on the interference of AI in online discourse. Minds Mach. online first https://doi.org/10.1007/s11023-024-09658-0 (2024)
24. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PLOS ONE. **14** (2019). https://doi.org/10.1371/journal.pone.0221152
25. U.S. President's executive order on the: safe, secure, and trustworthy development and use of artificial intelligence from 30th (2023). https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence (2023). Accessed 2 October 2024
26. EU building trust in human-centric AI (COM(2019)168:). (2019). https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence Accessed 17 April 2024
27. Huq, A.Z.: A right to a human decision. Va. L Rev. **106**, 611–688 (2020). https://virginialawreview.org/articles/right-human-decision/
28. EU ethics guidelines for trustworthy AI: (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Accessed 2 October 2024
29. EU Artificial Intelligence Act: (2024). https://artificialintelligenceact.eu/the-act/ Accessed 2 October 2024
30. Mökander, J., Axente, M., Casolari, F., Floridi, L.: Conformity assessments and post-market monitoring: a guide to the role of monitoring in the proposed European AI regulation. Minds Mach. **32**, 241–268 (2022). https://doi.org/10.1007/s11023-021-09577-4
31. Wagner, B.: Free expression? Dominant information intermediaries as arbiters of internet speech. In: Moore, M., Tambini, D. (eds.) Digital Dominance. The Power of Google, Amazon, Facebook, and Apple, pp. 219–240. Oxford University Press, New York (2018)
32. Proposal for a regulation of the European Parliament: and of the Council laying down harmonized rules on artificial intelligence (AI Act) and amending certain Union legislative acts (2024). https://artificialintelligenceact.eu/wp-content/uploads/2024/01/AIA-Final-Draft-21-January-2024.pdf Accessed 2 October 2024
33. Laux, J.: Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI Governance under the European Union AI Act. AI & Soc (2023). https://doi.org/10.1007/s00146-023-01777-z
34. Koulu, R.: Human control over automation: EU policy and AI ethics. EJLS. **12**, 9–46 (2020). https://cadmus.eui.eu/handle/1814/66992
35. Green, B.: The flaws of policies requiring human oversight of government algorithms. Comput. Law Sec Rev. **45** (2022). https://doi.org/10.1016/j.clsr.2022.105681
36. Schulz, W.: Regulating intermediaries to protect personality rights online—the case of the German NetzDG. In: Albers, M., Sarlet, I.W. (eds.) Personality and data Protection Rights on the Internet, pp. 289–306. Springer, Heidelberg (2022)
37. Meta Oversight Board Charter: (2024). https://www.oversightboard.com/wp-content/uploads/2024/03/OB_Charter_March_2024.pdf Accessed 2 October 2024
38. Douek, E.: What kind of Oversight Board have you given us? U. Chi. L Rev. Online, 05/11/2020 https://lawreview.uchicago.edu/online-archive/what-kind-oversight-board-have-you-given-us
39. Meta Oversight Board Bylaws: (2024). https://www.oversightboard.com/wp-content/uploads/2024/03/Oversight-Board-Bylaws.pdf Accessed 2 October 2024
40. Ang, P.H., Haristya, S.: The governance, legitimacy and efficacy of Facebook's Oversight Board: a model for global tech platforms? Emerg. Media, online first (2024). https://doi.org/10.1177/27523543241266860
41. Kaminski, M.E.: Binary governance: lessons from the GDPR's approach to algorithmic accountability. S Cal L Rev. **92**, 1529–1616 (2019).
42. Rebala, G., Ravi, A., Churiwala, S.: An Introduction to Machine Learning. Springer, Cham (2019)
43. Kaufmann, T., et al.: A survey of reinforcement learning from human feedback (2024). https://arxiv.org/pdf/2312.14925
44. McIntosh, T.R., et al.: The inadequacy of reinforcement learning from human feedback—radicalizing large language models via semantic vulnerabilities. IEEE Trans. Cogn. Dev. Syst. **16**, 1561–1574 (2024)
45. Douek, E.: Governing online speech: from posts-as-trumps to proportionality and probability. Colum L Rev. **121**, 759–834 (2021). https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/